Noname manuscript No.
(will be inserted by the editor)

A comparison of research data management platforms

Architecture, flexible metadata and interoperability

Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, Cristina Ribeiro

Received: date / Accepted: date

Abstract Research data management is rapidly becoming a regular concern for researchers, and institutions need to provide them with platforms to support data organization and preparation for publication. Some institutions have adopted institutional repositories as the basis for data deposit, whereas others are experimenting with richer environments for data description, in spite of the diversity of existing workflows. This paper is a synthetic overview of current platforms that can be used for data management purposes. Adopting a pragmatic view on data management, the paper focuses on solutions that can be adopted in the long-tail of science, where investments in tools and manpower are modest. First, a broad set of data management platforms is presented—some designed for institutional repositories and digital libraries—to select a short list of the more promising ones for data management. These platforms are compared considering their architecture, support for metadata, existing programming interfaces, as well as their search mechanisms and community acceptance. In this process, the stakeholders' requirements are also taken into account. The results show that there is still plenty of room for improvement, mainly regarding the specificity of data description in different domains, as well as the potential for integration of the data management platforms with existing research management tools. Nevertheless, depending on the context, some platforms can meet all or part of the stakeholders' requirements.

This paper is an extended version of a previously published comparative study. Please refer to the WCIST 2015 conference proceedings (doi: 10.1007/978-3-319-16486-1)

Ricardo Carvalho Amorim
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: ricardo.amorim3@gmail.com

João Aguiar Castro
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: joaoaguiarcastro@gmail.com

João Rocha da Silva
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: joaorosilva@gmail.com

Cristina Ribeiro
INESC TEC—Faculdade de Engenharia da Universidade do Porto
E-mail: mcr@fe.up.pt

1 Introduction

The number of published scholarly papers is steadily increasing, and there is a growing awareness of the importance, diversity and complexity of data generated in research contexts [25]. The management of these assets is currently a concern for both researchers and institutions who have to streamline scholarly communication, while keeping record of research contributions and ensuring the correct licensing of their contents [23, 18]. At the same time, academic institutions have new mandates, requiring data management activities to be carried out during the research projects, as a part of research grant contracts [14, 26]. These activities are invariably supported by software platforms, increasing the demand for such infrastructures.

This paper presents an overview of several prominent research data management platforms that can be put in place by an institution to support part of its research data management workflow. It starts by identifying a set of well known repositories that are currently being used for either publications or data management, discussing their use in several research in-

stitutions. Then, focus moves to their fitness to handle research data, namely their domain-specific metadata requirements and preservation guidelines. Implementation costs, architecture, interoperability, content dissemination capabilities, implemented search features and community acceptance are also taken into consideration. When faced with the many alternatives currently available, it can be difficult for institutions to choose a suitable platform to meet their specific requirements. Several comparative studies between existing solutions were already carried out in order to evaluate different aspects of each implementation, confirming that this is an issue with increasing importance [16,3,6]. This evaluation considers aspects relevant to the authors' ongoing work, focused on finding solutions to research data management, and takes into consideration their past experience in this field [33]. This experience has provided insights on specific, local needs that can influence the adoption of a platform and therefore the success in its deployment.

It is clear that the effort in creating metadata for research datasets is very different from what is required for research publications. While publications can be accurately described by librarians, good quality metadata for a dataset requires the contribution of the researchers involved in its production. Their knowledge of the domain is required to adequately document the dataset production context so that others can reuse it. Involving the researchers in the deposit stage is a challenge, as the investment in metadata production for data publication and sharing is typically higher than that required for the addition of notes that are only intended for their peers in a research group [7].

Moreover, the authors look at staging platforms, which are especially tailored to capture metadata records as they are produced, offering researchers an integrated environment for their management along with the data. As this is an area with several proposals in active development, EUDAT, which includes tools for data staging, and Dendro, a platform proposed for engaging researchers in data description, taking into account the need for data and metadata organisation will be contemplated.

Staging platforms are capable of exporting the enclosed datasets and metadata records to research data repositories. The platforms selected for the analysis in the sequel as candidates for use are considered as research data management repositories for datasets in the long tail of science, as they are designed with sharing and dissemination in mind. Together, staging platforms and research data repositories provide the tools to handle the stages of the research workflow. Long-term preservation imposes further requirements, and other tools may be necessary to satisfy them. However, as datasets become organised and described, their value and their potential for reuse will prompt further preservation actions.

2 From publications to data management

The growth in the number of research publications, combined with a strong drive towards open access policies [8,10], continue to foster the development of open-source platforms for managing bibliographic records. While data citation is not yet a widespread practice, the importance of citable datasets is growing. Until a culture of data citation is widely adopted, however, many research groups are opting to publish so-called "data papers", which are more easily citable than datasets. Data papers serve not only as a reference to datasets but also document their production context [9].

As data management becomes an increasingly important part of the research workflow [24], solutions designed for managing research data are being actively developed by both open-source communities and data management-related companies. As with institutional repositories, many of their design and development challenges have to do with description and long-term preservation of research data. There are, however, at least two fundamental differences between publications and datasets: the latter are often purely numeric, making it very hard to derive any type of metadata by simply looking at their contents; also, datasets require detailed, domain-specific descriptions to be correctly interpreted. Metadata requirements can also vary greatly from domain to domain, requiring repository data models to be flexible enough to adequately represent these records [35]. The effort invested in adequate dataset description is worthwhile, since it has been shown that research publications that provide access to their base data consistently yield higher citation rates than those that do not [27].

As these repositories deal with a reasonably small set of managed formats for deposit, several reference models, such as the OAIS (Open Archival Information System) [12] are currently in use to ensure preservation and to promote metadata interchange and dissemination. Besides capturing the available metadata during the ingestion process, data repositories often distribute this information to other instances, improving the publications' visibility through specialised research search engines or repository indexers. While the former focus on querying each repository for exposed contents, the latter help users find data repositories that match their needs—such as repositories from a specific domain or storing data from a specific community. Governmental

institutions are also promoting the disclosure of open data to improve citizen commitment and government transparency, and this motivates the use of data management platforms in this context.

2.1 An overview on existing repositories

While depositing and accessing publications from different domains is already possible in most institutions, ensuring the same level of accessibility to data resources is still challenging, and different solutions are being experimented to expose and share data in some communities. Addressing this issue, we synthesize a preliminary classification of these solutions according to their specific purpose: they are either targeting staging, early research activities or managing deposited datasets and making them available to the community.

Table 1 identifies features of the selected platforms that may render them convenient for data management. To build the table, the authors resorted to the documentation of the platforms, and to basic experiments with demonstration instances, whenever available. In the first column, under "Registered repositories", is the number of running instances of each platform, according to the OpenDOAR platform as of mid-October 2015.

In the analysis, five evaluation criteria that can be relevant for an institution to make a coarse-grained assessment of the solutions are considered. Some existing tools were excluded from this first analysis, mainly because some of their characteristics place them outside of the scope of this work. This is the case of platforms specifically targeting research publications (and that cannot be easily modified for managing data), and heavy-weight platforms targeted at long-term preservation. Also excluded were those that, from a technical point of view, do not comply with desirable requirements for this domain such as adopting an open-source approach, or providing access to their features via comprehensive APIs.

By comparing the number of existing installations, it is natural to assume that a large number of instances for a platform is a good indication of the existence of support for its implementation. Repositories such as DSpace are widely used among institutions to manage publications. Therefore, institutions using DSpace to manage publications can use their support for the platform to expand or replicate the repository and meet additional requirements.

It is important to mention that some repositories do not implement interfaces with existing repository indexers, and this may cause the OpenDOAR statistics to show a value lower than the actual number of existing installations. Moreover, services provided by EUDAT, Figshare and Zenodo, for instance, consist of a single installation that receives all the deposited data, rather than a distributed array of manageable installations.

Government-supported platforms such as CKAN are currently being used as part of the open government initiatives in several countries, allowing the disclosure of data related to sensitive issues such as budget execution, and their aim is to vouch for transparency and credibility towards tax payers [21,20]. Although not specifically tailored to meet research data management requirements, these data-focused repositories also count with an increasing number of instances supporting complex research data management workflows [38], even at universities[1].

Access to the source code can also be a valuable criterion for selecting a platform, primarily to avoid vendor lock-in, which is usually associated with commercial software or other provided services. Vendor lock-in is undesirable from a preservation point of view as it places the maintenance of the platform (and consequently the data stored inside) in the hands of a single vendor, that may not be able to provide support indefinitely. The availability of the a platform's source code also allows additional modifications to be carried out in order to create customized workflows—examples include improved metadata capabilities and data browsing functionalities. Commercial solutions such as ContentDM may incur high costs for the subscription fees, which can make them cost-prohibitive for non-profit organizations or small research institutions. In some cases only a small portion of the source code for the entire solution is actually available to the public. This is the case with EUDAT, where only the B2Share module is currently open[2]—the remaining modules are unavailable to date.

From an integration point of view, the existence of an API can allow for further development and help with the repository maintenance, as the software ages. Solutions that do not, at least partially, comply with this requirement, may hinder the integration with external platforms to improve the visibility of existing contents. The lack of an API creates a barrier to the development of tools to support a platform in specific environments, such as laboratories that frequently produce data to be directly deposited and disclosed. Finally, regarding long-term preservation, some platforms fail to provide unique identifiers for the resources upon deposit, making persistent references to data and data citation in publications hard.

[1] `http://ckan.org/2013/11/28/ckan4rdm-st-andrews/`

[2] Source code repository for B2Share is hosted via GitHub at `https://github.com/EUDAT-B2SHARE/b2share`

Table 1: Limitations of the identified repository solutions. **Source:** ▽OpenDOAR platform △Corresponding website. †Only available through additional plug-ins. *Only partially.

	Registered repositories▽	Closed source	No API	No unique identifiers	Complex installation or setup	No OAI-PMH compliance
CKAN	139△			×†		×*
ContentDM	53	×				
Dataverse	2					
Digital Commons	141	×	×			
DSpace	1305					
ePrints	407			×†		
EUDAT	—	×*				
Fedora	41				×	
Figshare	—	×				
Greenstone	51		×	×	×	
Invenio	20					
Omeka	4			×		×†
SciELO	18	×				
WEKO	40			No data		
Zenodo	—					

Support for flexible research workflows makes some repository solutions attractive to smaller institutions looking for solutions to implement their data management workflows. Both DSpace and ePrints, for instance, are quite common as institutional repositories to manage publications, as they offer broad compatibility with the harvesting protocol OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) [22] and with preservation guidelines according to the OAIS model. OAIS requires the existence of different packages with specific purposes, namely SIP (Submission Information Package), AIP (Archival Information Package) and DIP (Dissemination Information Package). The OAIS reference model defines SIP as a representation of packaged items to be deposited in the repository. AIP, on the other hand, represents the packaged digital objects within the OAIS-compliant system, and DIP holds one or several digital artifacts and their representation information, in such a format that can be interpreted by potential users.

2.2 Stakeholders in research data management

Several stakeholders are involved in dataset description throughout the data management workflow, playing an important part in their management and dissemination [24,7]. These stakeholders—*researchers*, *research institutions*, *curators*, *harvesters*, and *developers*—play a governing role in defining the main requirements of a data repository for the management of research outputs. As key metadata providers, *researchers* are responsible for the description of research data. They are not necessarily knowledgeable in data management practices, but can provide domain-specific, more or less formal descriptions to complement generic metadata. This captures the essential data production context, making it possible for other researchers to reuse the data [7]. As data creators, researchers can play a central role in data deposit by selecting appropriate file formats for their datasets, preparing their structure and packaging them appropriately [15]. *Institutions* are also motivated to have their data recognized and preserved according to the requirements of funding institutions [17, 26]. In this regard, institutions value metadata in compliance to standards, which make data ready for inclusion in networked environments, therefore increasing their visibility. To make sure that this context is correctly passed, along with the data, to the preservation stage, *curators* are mainly interested in maintaining data quality and integrity over time. Usually, curators are information experts, so it is expected that their close collaboration with researchers can result in both detailed and compliant metadata records.

Considering data dissemination and reuse, *harvesters* can be either individuals looking for specific data

or services which index the content of several repositories. These services can make particularly good use of established protocols, such as the OAI-PMH, to retrieve metadata from different sources and create an interface to expose the indexed resources. Finally, contributing to the improvement and expansion of these repositories over time, *developers* are concerned with the underlying technologies, an also in having extensive APIs to promote integration with other tools.

3 Scope of the analysis

The stakeholders in the data management workflow can greatly influence whether research data is reused. The selection of platforms in the analysis acknowledges their role, as well as the importance of the adoption of community standards to help with data description and management in the long run.

For this comparison, data management platforms with instances running at both research and government institutions have been considered, namely DSpace, CKAN, Zenodo, Figshare, ePrints, Fedora and EUDAT. If the long-term preservation of research assets is an important requirement of the stakeholders in question, other alternatives such as RODA [30] and Archivematica may also be considered strong candidates, since they implement comprehensive preservation guidelines not only for the digital objects themselves but also for their whole life cycle and associated processes. On one hand, these platforms have a strong concern with long-term preservation by strictly following existing standards such as OAIS, PREMIS or METS, which cover the different stages of a long-term preservation workflow. On the other hand, such solutions are usually harder to install and maintain by institutions in the so-called long tail of science—institutions that create large numbers of small datasets, though do not possess the necessary financial resources and preservation expertise to support a complete preservation workflow [18].

The Fedora framework[3] is used by some institutions, and is also under active development, with the recent release of Fedora 4. The fact that it is designed as a framework to be fully customized and instantiated, instead of being a "turnkey" solution, places Fedora in a different level, that can not be directly compared with other solutions. Two open-source examples of Fedora's implementations are Hydra[4] and Islandora[5]. Both are open-source, capable of handling research workflows, and use the best-practices approach already implemented in the core Fedora framework. Although these are not present in the comparison table, this section will also consider their strengths, when compared to the other platforms.

An overview of the previously identified stakeholders led to the selection of two important dimensions for the assessment of the platform features: their architecture and their metadata and dissemination capabilities. The former includes aspects such as how they are deployed into a production environment, the locations where they keep their data, whether their source code is available, and other aspects that are related to the compliance with preservation best practices. The latter focuses on how resource-related metadata is handled and the level of compliance of these records with established standards and exchange protocols. Other important aspects are their adoption within the research communities and the availability of support for extensions. Table 2 shows an overview of the results of our evaluation.

4 Platform comparison

Based on the selection of the evaluation scope, this section addresses the comparison of the platforms according to key features that can help in the selection of a platform for data management. Table 2 groups these features in two categories: (i) Architecture, for structural-related characteristics; and (ii) Metadata and dissemination, for those related to flexible description and interoperability. This analysis is guided by the use cases in the research data management environment.

4.1 Architecture

Regarding the architecture of the platforms, several aspects are considered. From the point of view of a research institution, a quick and simple deployment of the selected platform is an important aspect. There are two main scenarios: the institution can either outsource an external service or install and customize its own repository, supporting the infrastructure maintenance costs. Contracting a service provided by a dedicated company such as Figshare or Zenodo delegates platform maintenance for a fee. The service-based approach may not be viable in some scenarios, as some researchers or institutions may be reluctant to deposit their data in a platform outside their control [11]. DSpace, ePrints, CKAN or any Fedora-based solution can be installed and run completely under the control of the research institution and therefore offer a better control over the stored data. As open-source solutions, they also have

[3] http://www.fedora-commons.org/
[4] http://projecthydra.org/
[5] http://islandora.ca/

Table 2: Comparison of the selected research data management platforms

	Feature	DSpace	CKAN	Figshare	Zenodo	ePrints	EUDAT
Architecture	**Deployment**	Installation package or service	Installation package	Service	Service	Installation package or service	Service
	Storage Location	Local or remote	Local or remote	Remote	Remote	Local or remote	Remote
	Maintenance costs	Infrastructure management	Infrastructure management	Monthly fee	Monthly fee	Infrastructure management	Monthly fee
	Open Source	✓	✓	×	×	✓	×
	Customization	✓	✓	×	Community policies	✓	×
	Internationalization support	✓	✓	×	×	✓	×
	Embargo	✓	Private Storage	Private Storage	✓	✓	✓
	Content versioning	×	✓	×	×	✓	✓
	Pre-reserving DOI	✓	×	✓	✓	✓	✓
Metadata & Dissemination	**Exporting schemas**	Any pre-loaded schemas	None	DC	DC, MARCXML	DC, METS, MODS, DIDL	DC, MARC, MARCXML
	Schema flexibility	Flexible	Flexible	Fixed	Fixed	Fixed	Flexible
	Validation	✓	×	×	✓	✓	✓
	Versioning	×	✓	×	×	✓	✓
	OAI-PMH	✓	×	✓	✓	✓	✓
	Record license specification	✓	✓	✓	✓	✓	✓

several supporters[6] that contribute to their expansion with additional plugins or extensions to meet specific requirements. DSpace, CKAN and Zenodo allow a certain degree of customization to satisfy the needs of their users: while Zenodo allows parametrization settings such as community-level policies, CKAN, DSpace and Fedora—as open source solutions—can be further customized, with improvements ranging from small interface changes to the development of new data visualization plugins [33,34]. Due to its complex architecture, DSpace may require a higher level of expertise when dealing with custom features. However, its larger supporting community may help tackling such barriers. The same applies to Fedora as it requires the research institution to choose among different technologies to design and implement the end-user interface, which can exclude it as an option if limited time or budget restrictions apply. A positive aspect in all packaged platforms is that they provide easy internationalization support. The Zenodo and Figshare services are available in English only, as well as the majority of EUDAT's interfaces—an exception is its B2Share module, which is built on the Invenio platform, which already has internationalization features.

A collaborative environment for teams and groups to manage the deposited resources is becoming increasingly important in the research workflows of many institutions. In this regard, both CKAN and Zenodo provide collaborative tools and allow users to fully manage their group members and policies. ePrints and Dspace are not designed to support real-time collaborative environments where researchers can produce data and describe them incrementally, so these platforms can be less suited to support dynamic data production environments. Adopting a dynamic approach to data management, tasks can be made easier for the researchers, and motivate them to use the data management platform as part of their daily research activities, while they are working on the data. Otherwise, researchers may only consider depositing data in the platform after datasets are finished—no longer in active gathering or processing—and this is likely to reduce the number of datasets that get into the deposit phase. Moreover,

[6] `http://ckan.org/instances/`
`http://registry.duraspace.org/registry/dspace`

different researchers may have a different approach to dataset structure and description, and this will cause difficulties to the workflows that rely solely on deposit. EUDAT provides a collaborative environment by integrating file management and sharing into the research workflow via a desktop application. This application can automatically synchronize files to one of the environment's modules (B2Drop). After the files are uploaded, they can be used for computation in B2Stage or shared in B2Share to major portals in several research areas. They can also be available for search in B2Find, the repository of the EUDAT environment designed to e an aggregator for metadata on research datasets. EUDAT's B2Share service is built on the Invenio data management platform. This platform is flexible, available under an open-source license, and compatible with several metadata representations, while still providing a complete API. However, it could be hard to manage and possibly decommission an Invenio platform in the future, since its underlying relational model is complex and very tightly connected to the platform's code [35].

The control over data release dates can also be a concern for researchers. DSpace, ePrints, Zenodo and EUDAT allow users to specify embargo periods; data is made available to the community after they expire. CKAN and Figshare have options for private storage, to let researchers control the data publication mode.

4.2 Metadata: a key for preservation

Research data can benefit from domain-level metadata to contextualize their production [37]. While the evaluated platforms have different description requirements upon deposit, most of them lack the support for domain-specific metadata schemas. In this regard DSpace is an exception, with its ability to use multiple schemas that can be set up by a system administrator. The same happens with Islandora, which uses the support for descriptive metadata available in Fedora, allowing the creation of tailored metadata forms, if the corresponding plugin is installed. This is a solution for the requirement of providing research data with domain-level metadata, a matter that is still to be addressed by several other platforms. Both Zenodo and Figshare can export records that comply with established metadata schemas (Dublin Core and MARC-XML, and Dublin Core, respectively). DSpace goes further by exporting DIPs that include METS metadata records, thus enabling the ingestion of these packages into a long-term preservation workflow. Although CKAN metadata records do not follow any standard schema, the platform allows the inclusion of a dictionary of key-value pairs that can be used, for instance, to record domain-specific metadata as a complement to generic metadata descriptions. Neither of these platforms natively supports collaborative validation stages where curators and researchers enforce the correct data and metadata structure, although Zenodo allows the users to create a highly curated area within communities, as highlighted in the "validation" feature in Table 2. If the policy of a particular community specifies manual validation, every deposit will have to be validated by the community curator. EUDAT does not support domain-dependent metadata, however it can gather different sets of descriptors when depositing to different projects using B2Share. For example, when the user performing the deposit chooses GBIF (a biodiversity infrastructure) as the target project for the new dataset, some predefined, biodiversity-related descriptors become available to be filled in as a complement to the generic ones. These domain-specific descriptors can greatly improve generic descriptions. Datasets originate from very specific research domains, thus requiring specific descriptions to be correctly interpreted by potential users.

Tracking content changes is also an important issue in data management, as datasets are often versioned and dynamic. CKAN provides an auditing trail of each deposited dataset by showing all changes made to it since its deposit. EUDAT deals with the problem of metadata auditing in the same way, because its dataset search and retrieval engine, B2Find, is based on CKAN technology[7], and can therefore provide the same auditing trail interface.

4.3 Interoperability and dissemination

Exposing repository contents to other research platforms can improve both data visibility and reuse [24]. All of the evaluated platforms allow the development of external clients and tools as they already provide their own APIs for exposing metadata records to the outside community, with some differences regarding standards compliance. In this matter, only CKAN is not natively compliant with OAI-PMH. This is a widely-used protocol that promotes interoperability between repositories while also streamlining data dissemination, and is a valuable resource for harvesters to index the contents of the repository [22,13]. As an initiative originally designed for government data, it is understandable that CKAN is missing this compliance, although it can leave institutions reluctant to its adoption as they can also have interest in getting their datasets cited by the community.

[7] Please refer to `http://eudat.eu/sites/default/files/DaanBroeder.pdf`

Table 3: Key advantages of the evaluated repository platforms

Platform	Key advantages
Figshare	– Gives credit to authors through citations and references – Can export reference to Mendeley, DataCite, RefWorks, Endnote, NLM and ReferenceManager – Records statistics related to citations and shares – Does not require any maintenance
Zenodo	– Allows creating communities to validate submissions – Supports Dublin Core, MARC and MARCXML for metadata exporting – Can export references to BibTeX, DataCite, DC, EndNote, NLM, RefWorks – Complies with OAI-PMH for data dissemination – Does not require any maintenance – Includes metadata records in the searchable fields
CKAN	– Is open-source and widely supported by the developer community – Features extensive and comprehensive documentation – Allows deep customization of its features – Can be fully under institutions control – Supports unrestricted (non standards-compliant) metadata – Has faceted search with fuzzy-matching – Records datasets change logs and versioning information
DSpace	– Can comply with domain-level metadata schemas – Is open-source and has a wide supporting community – Has an extensive, community maintained documentation – Can be fully under institutions control – Structured metadata representation – Compliant with OAI-PMH
ePrints	– Can maintain records of changes in preservation metadata records – Compliant with OAI-PMH – Compliant with SWORD for multiple deposit
EUDAT	– Modular approach that provides a variety of services to match local needs – Strong support form European agencies – Integration of several open-source platforms (CKAN, Invenio) – End-to-end workflow for research data management – Majority of features are available for free to european researchers

It is interesting not only to evaluate platforms according to the ease of discovery by machines, but also to see how easily humans can find a dataset there. All three platforms possess free-text search capabilities, indexing the metadata in dataset records for retrieval purposes. All analyzed platforms provide an "advanced" search feature that is in practice a faceted search. Depending on the platform, users can restrict the results to smaller sets, for instance from a domain such as Engineering. This search feature makes it easier for researchers to find the datasets that are from relevant domains and belong to specific collections or similar dataset categories (the concept varies between platforms as they have different organizational structures). ePrints, for instance, allows search on the metadata records, includes boolean operators to refine the results as well as full text search for some of the compatible data formats, provided the appropriate plugins are installed. When considering the involved technologies, DSpace stands out as it natively uses Apache Lucene as a search engine which competes with the Xapian[8] engine used in ePrints, to sort results by relevance.

[8] `http://xapian.org/`

4.4 Platform adoption

As most recent platforms, all the repositories depend on a community of developers to maintain and improve their features. Looking for successful case studies, it is important to assess their impact and comprehensiveness. CKAN has several success cases with government data which are made available to the community, although missing other scenarios related to the management and disclosure of research data. Figshare, Zenodo and DSpace have research data as their focus. In active use since 2002, DSpace is well known among institutions and researchers for its capabilities to deal with research publications and, more recently, also to handle research data. DSpace benefits from a dominant position in institutional repositories and the existence of such instances can favour its adoption for dataset management. Zenodo is a solution for the long tail of science supported by CERN laboratories, and is regarded as an environment to bring research outputs to an appropriate digital archive for preservation. It is therefore also a strong use case, with researchers from many fields already using it.

5 Data staging platforms

Most of the analyzed solutions target data repositories, i.e. the end of the research workflow. They are designed to hold and manage research data outputs after the data production is concluded and the results of their analysis are published. As a consequence, there is an overall lack of support for capturing data during the earlier stages of research activities.

Introducing data management—and metadata production particularly—at an early stage in the research workflow increases the chances of a dataset reaching the final stage of this workflow, when it is kept in a long-term preservation environment. The introduction of data deposit and description earlier in the research workflow means that descriptions will already be partially done by the end of data gathering. Also, more detailed and overall better metadata records can be created in this way, since the data creation context is still present. Researchers can also reap immediate benefits from their data description, as described datasets can more easily be shared among the members of their research group or with external partners.

Data gathering is often a collaborative process, so it makes sense to make metadata production collaborative as well. These requirements have been identified by several research and data management institutions, who have implemented integrated solutions for researchers to manage data not only when it is created, but also throughout the entire research workflow.

Researchers are not data management experts, so they need effective tools that allow them to produce adequate standards-compliant metadata records without having to learn about those standards. Thus, an important characteristic of an effective solution for collaborative data management is its ease of use by non-experts. If these solutions are easy to use and provide both immediate and long-term added value for researchers, they are more likely to be adopted as part of the daily research work. Gradually, this would counteract the idea that data management is a time-consuming process performed only due to policies enforced by funding institutions, or motivated by uncertain and long-term rewards such as the possibility of others citing the datasets.

5.1 Data management as a routine task

There have been important advancements towards the incorporation of data management practices in the day-to-day activities of researchers.

In the UK, the DataFlow project [19] was built to provide researchers with an integrated data management workflow to allow them to store and describe their data safely and easily. The project implemented two components: DataStage and DataBank. DataStage allows researchers standards-based (CIFS, SFTP, SSH, WebDAV) access to shared data storage areas protected by automated backups, as well as a web interface that researchers can use to add metadata to the files that they deposit. The shared storage is accessible from the computers used for their work through a mapped drive. When researchers are ready to deposit a dataset, they can package it as a ZIP file and send it to DataBank via a SWORD endpoint. DataBank is a repository platform that, besides supporting ingestion via the SWORD v2 protocol, supports DOI registration via DataCite, version control, specification of embargo periods and OAI-PMH compliance to foster the dissemination of data. File format-related operations—such as correct identification of the format for a file—are handled by existing tools such as JHOVE and DROID [5].

datorium is a platform for the description and sharing of research data from the social sciences. Realizing the increasing requirements for base data as supplementary material to research publications, its goal is to provide an easy to use platform for researchers to perform autonomous description of their datasets. Metadata is, like other platforms, limited to Dublin Core, complemented in this case with some elements taken from GESIS Data Catalogue DBK [1].

MaDAM [28] is a web-based data management system targeted at the management of research data in research groups. It provides a user-friendly file explorer, as well as an editor for adding metadata to the entities in the folder structure. The descriptors that can be added to a metadata record are fixed and general-purpose, such as "Name", "Creator" or "Comments". The platform also has an "Archive" function that allows users to send a dataset to eScholar, the University of Manchester's preservation and dissemination repository[9].

DASH[10], a data management platform in use at the University of California, incorporates two previous tools: DataUP[11] and DataShare[12]. It does not currently support interoperability protocols for deposit or dissemination of datasets such as OAI-PMH or SWORD, which leaves it outside of the present comparison. However, it is an open-source project, and its modules are currently available[13]. It also provides an easy to use interface, indexing by scholarly engines, data identifica-

[9] http://www.escholar.manchester.ac.uk/
[10] https://dash.library.ucsc.edu/
[11] http://dataup.cdlib.org
[12] http://datashare.ucsf.edu/xtf/search
[13] http://cdluc3.github.io/dash/

tion via DOI and integration with Merritt, an in-house developed long-term repository[14].

HAL is a platform for the deposit, description and dissemination of research datasets. It provides a wikipedia plugin to modify the layout of Wikipedia pages and directly include links to datasets. This can help researchers find data in Wikipedia pages. The metadata that can be added to each dataset is limited to a set of generic, fixed descriptors, whose values can be derived from the content of relevant Wikipedia pages [29].

As a pan-european effort for the creation of an integrated research data management environment, EUDAT also includes a file sharing module called B2Drop. It provides researchers with 20GB of storage for free, and is integrated with other modules for dataset sharing and staging, including some computational processing on the stored data.

Several interesting concepts have been recently presented as part of an integrated vision for the management of research data within research groups. Some core concepts currently found in social networks can be applied to research data management, making it a natural part of the daily activities of researchers [4]. They include a timeline of changes over resources under the group's control, comments that are linked to those changes, external sharing controlled by the elements of the research group and the ability to track the interactions of external entities with the dataset (such as citations and "likes"). In this view, researchers are able to browse datasets deposited by group members as they are produced, and also run workflows over that data. The continuous recording of both data and the translation steps that allow a dataset to be derived from others is a very interesting concept not only from a preservation point of view, but also in scientific terms, as it safeguards the reproducibility of research findings.

5.2 Dendro

UPBox and DataNotes where designed an implemented at the University of Porto as coupled solutions to provide users with an integrated data management environment [31]. UPBox was designed to provide researchers with a shared data storage environment, fully under their research institution control and complemented by an easy-to-use REST API to allow its integration with multiple services. DataNotes was a modified version of Semantic MediaWiki, designed to work with UPBox, allowing researchers to produce wiki-formatted pages describing the files and folders that they had previously sent to the data storage environment. The generated metadata would use descriptors from diverse ontologies from multiple domains and could be exported as RDF records.

The lessons learned during the implementation of these two solutions and through the ongoing analysis of requirements in research groups, led to the development of Dendro. Dendro is a single solution targeted at improving the overall availability and quality of research data. It aims at engaging researchers in the management and description of their data, focusing on metadata recording at the early stages of the research workflow [32,36]. Dendro is a fully open-source environment (solution and dependencies) that combines an easy to use file manager (similar to Dropboxb) with the collaborative capabilities of a semantic wiki for the production of semantic metadata records. The solution aims at the description of datasets from different research domains through an extensible, triple store-based data model [35]. Curators can expand the platform's data model by loading ontologies that specify domain-specific or generic metadata descriptors that can then be used by researchers in their projects. These ontologies can be designed using tools such as Protégé[15], allowing curators with no programming background to extend the platform's data model. Dendro is designed primarily as a staging environment for dataset description. Ideally, as research publications are written, associated datasets (already described at this point) are packaged and sent to a research data repository, where they go through the deposit workflows. In the end, the process can be made fast enough to enable researchers to cite the datasets in the publication itself as supporting data.

Dendro focuses on interoperability to make the deposit process as easy as possible for researchers. It can be integrated with all the repository platforms surveyed in this paper, while its extensive API makes it easy to integrate with external systems. LabTablet, an electronic laboratory notebook designed to help researchers gather metadata in experimental contexts, is an example of a successful integration scenario. It allows researchers to generate metadata records using the mobile device's on-board sensors, which are then represented using established metadata schemas (e.g. Dublin Core) and uploaded to a Dendro instance for collaborative editing [2].

[14] `http://guides.library.ucsc.edu/datamanagement/publish`

[15] Available at `http://protege.stanford.edu/`

6 Conclusion

The evaluation showed that it can be hard to select a platform without first performing a careful study of the requirements of all stakeholders. The main positive aspects of the platforms considered here are summarized in Table 3. Both CKAN and DSpace's open-source licenses that allow them to be updated and customized, while keeping the core functionalities intact, are highlighted.

Although CKAN is mainly used by governmental institutions to disclose their data, its features and the extensive API making it also possible to use this repository to manage research data, making use of its key-value dictionary to store any domain-level descriptors. This feature does not however strictly enforce a metadata schema. Curators may favor DSpace though, since it enables system administrators to parametrize additional metadata schemas that can be used to describe resources. These will in turn be used to capture richer domain-specific features that may prove valuable for data reuse.

Researchers need to comply with funding agency requirements, so they may favour easy deposit combined with easy data citation. Zenodo and Figshare provide ways to assign a permanent link and a DOI, even if the actual dataset is under embargo at the time of first citation. This will require a direct contact between the data creator and the potential reuser before access can be provided. Both these platforms are aimed at the direct involvement of researchers in the publication of their data, as they streamline the upload and description processes, though they do not provide support for domain-specific metadata descriptors.

A very important factor to consider is also the control over where the data is stored. Some institutions may want the servers where data is stored under their control, and to directly manage their research assets. Platforms such as DSpace and CKAN, that can be installed in an institutional server instead of relying on external storage provided by contracted services are appropriate for this.

The evaluation of research data repositories can take into account other features besides those considered in this analysis, namely their acceptance within specific research communities and their usability. The paper has focused on repositories as final locations for research data to be deposited and not as a replacement for the tools that researchers already use to manage their data—such as file sharing environments or more complex e-science platforms. The authors consider that these solutions should be compared to other collaborative solutions such as Dendro, a research data management solution currently under development. In this regard, it can be argued that flexible, customizable solutions such as Dendro can meet the needs of research institutions in terms of staging, temporary platforms to help with research data management and description. This should, of course, be done while taking into consideration available metadata standards that can contribute to overall better conditions for long-term preservation [36].

EUDAT features the integration of open-source established solutions (such as CKAN and Invenio) to support a comprehensive data management workflow. The platform is backed by several prominent institutions and promises to deliver an European data management environment to support research. Areas for improvement in this project include metadata production and collaboration. For example, limited domain-specific descriptors are available depending on the portal to which the dataset is being sent, instead of a fully flexible and expansible metadata model that depends on the research domain, such as the one in Dendro [35,36]). Collaboration challenges include the implementation of social-network based concepts for real-time collaboration [4].

Considering small institutions that somehow struggle to contract a dedicated service for data management purposes, having a wide community supporting the development of a stand-alone platform can be a valuable asset. In this regard, CKAN may have an advantage over the remaining alternatives, as several governmental institutions are already converging to this platform for data publishing.

Acknowledgements

This work is supported by the project NORTE-07-0124-FEDER000059, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). João Rocha da Silva is also supported by research grant SFRH/BD/77092/2011, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

1. Alam, A.W., Müller, S., Schumann, N.: datorium : Sharing platform for social science data. In: Proceedings of the 14th International Symposium on Information Science (ISI 2015), pp. 244–249 (2015)

2. Amorim, R.C., Castro, J.A., Rocha da Silva, J., Ribeiro, C.: Labtablet: Semantic metadata collection on a multi-domain laboratory notebook. Springer Communications in Computer and Information Science **478**, 193–205 (2014)
3. Armbruster, C., Romary, L.: Comparing repository types: challenges and barriers for subject-based repositories, research repositories, national repository systems and institutional repositories in serving scholarly communication. International Journal of Digital Library Systems **1**(4) (2010)
4. Assante, M., Candela, L., Castelli, D., Manghi, P., Pagano, P.: Science 2.0 repositories: Time for a change in scholarly communication. D-Lib Magazine **21**(1) (2015)
5. Ball, A.: Tools for research data management. Tech. rep., University of Bath, Bath, UK (2012)
6. Bankier, J.: Institutional repository software comparison. UNESCO Communication and Information **33** (2014)
7. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology **63**(6), 1059–1078 (2012)
8. Burns, C.S., Lana, A., Budd, J.: Institutional repositories: exploration of costs and value. D-Lib Magazine **19**(1), 1 (2013)
9. Candela, L., Castelli, D., Manghi, P., Tani, A.: Data journals: A survey. International Review of Research in Open and Distance Learning **66**, 1747–1762 (2015)
10. Coles, S.J., Frey, J.G., Bird, C.L., Whitby, R.J., Day, A.E.: First steps towards semantic descriptions of electronic laboratory notebook records. Journal of Cheminformatics **5**, 1–10 (2013)
11. Corti, L., Van den Eynden, V., Bishop, L., Woollard, M.: Managing and sharing research data: A guide to good practice. Records Management Journal **24**(3), 252–253 (2014)
12. Council of the Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS). Tech. Rep. January (2002)
13. Devarakonda, R., Palanisamy, G.: Data sharing and retrieval using OAI-PMH. Earth Science Informatics **4**(1), 1–5 (2011)
14. European Commission: Guidelines on open access to scientific publications and research data in horizon 2020. Tech. Rep. December (2013)
15. Van den Eynden, V., Corti, L., Bishop, L., Horton, L.: Managing and Sharing Data, 3 edn. UK Data Archive University of Essex (2011)
16. Fay, E.: Repository software comparison: building digital library infrastructure at LSE. Ariadne **64**(2009), 1–11 (2010)
17. Green, A., Macdonald, S., Rice, R.: Policy-making for Research Data in Repositories: A Guide. London: JISC funded DISC-UK Share Project (2009)
18. Heidorn, P.: Shedding light on the dark data in the long tail of science. Library Trends **57**(2), 280–299 (2008)
19. Hodson, S.: ADMIRAL: A Data Management Infrastructure for Research Activities in the Life sciences. Tech. rep., University of Oxford (2011)
20. Hoxha, J., Brahaj, A.: Open government data on the web: A semantic approach. In: International Conference on Emerging Intelligent Data and Web Technologies, pp. 107–113 (2011)
21. Kučera, J., Chlapek, D., Mynarz, J.: Czech CKAN repository as case study in public sector data cataloging. Systémová Integrace **19**(2), 95–107 (2012)
22. Lagoze, C., Sompel, H.V.D., Nelson, M., Warner, S.: The Open Archives Initiative Protocol for Metadata Harvesting. Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries (2001)
23. Lynch, C.A.: Institutional repositories: essential infrastructure for scholarship in the digital age. portal: Libraries and the Academy **3**(2), 327–336 (2003)
24. Lyon, L.: Dealing with Data: Roles, Rights, Responsibilities and Relationships. Tech. rep., UKOLN, University of Bath (2007)
25. McNutt, M.: Improving scientific communication. Science **342**(6154), 13 (2013)
26. National Science Foundation: Grants.gov Application Guide: A Guide for Preparation and Submission of National Science Foundation Applications via Grants.gov. Tech. rep. (2011)
27. Piwowar, H.A., Vision, T.J.: Data reuse and the open data citation advantage. PeerJ **1**, e175 (2013)
28. Poschen, M., Finch, J., Procter, R., Goff, M., McDerby, M., Collins, S., Besson, J., Beard, L., Grahame, T.: Development of a Pilot Data Management Infrastructure for Biomedical Researchers at University of Manchester—Approach, Findings, Challenges and Outlook of the MaDAM Project. International Journal of Digital Curation **7**, 110–122 (2012)
29. Rafes, K., Germain, C.: A platform for scientific data sharing. In: BDA2015 Bases de Donées Avancées (2015)
30. Ramalho, J.C., Ferreira, M., Faria, L., Castro, R., Barbedo, F., Corujo, L.: RODA and CRiB a service-oriented digital repository. iPres Conference Proceedings (2008)
31. Rocha da Silva, J., Barbosa, J., Gouveia, M., Correia Lopes, J., Ribeiro, C.: UPBox and DataNotes: a collaborative data management environment for the long tail of research data. iPres Conference Proceedings (2013)
32. Rocha da Silva, J., Castro, J.A., Ribeiro, C., Correia Lopes, J.: Dendro: collaborative research data management built on linked open data (2014)
33. Rocha da Silva, J., Ribeiro, C., Correia Lopes, J.: UPData—A Data Curation Experiment at U.Porto using DSpace. In: iPres Conference Proceedings, pp. 224–227 (2011)
34. Rocha da Silva, J., Ribeiro, C., Correia Lopes, J.: Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. In: iPres Conference Proceedings, pp. 105–108 (2012)
35. Rocha da Silva, J., Ribeiro, C., Correia Lopes, J.: Ontology-based multi-domain metadata for research data management using triple stores. In: Proceedings of the 18th International Database Engineering & Applications Symposium (2014)
36. Rocha da Silva, J., Ribeiro, C., Correia Lopes, J.: The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment. iPres Conference Proceedings (2014)
37. Willis, C., Greenberg, J., White, H.: Analysis and Synthesis of Metadata Goals for Scientific Data. Journal of the Association for Information Science and Technology **63**(8), 1505–1520 (2012)
38. Winn, J.: Open data and the academy: An evaluation of CKAN for research data management. International Association dor Social Science Information Services and Technology (2013)