IDCC17 | *Practice Paper*

Integration of an active research data system with a data repository to streamline the research data lifecycle: Pure-NOMAD case study

Simone Ivan Conte
University of St Andrews

Federica Fina
University of St Andrews

Michalis Psalios
University of St Andrews

Shyam Reyal
University of St Andrews

Tomas Lebl
University of St Andrews

Anna Clements
University of St Andrews

Abstract

Research funders have introduced requirements that expect researchers to properly manage and publicly share their research data, and expect institutions to put in place services to support researchers in meeting these requirements. So far the general focus of these services and systems has been on addressing the final stages of the research data lifecycle (archive, share and re-use), rather than stages related to the active phase of the cycle (collect/create and analyse). As a result, full integration of active data management systems with data repositories is not yet the norm, making the streamlined transition of data from an active to a published and archived status an important challenge. In this paper we present the integration between an active data management system developed in-house (NOMAD) and Elsevier's Pure data repository used at our institution with the aim of offering a simple workflow to facilitate and promote the data deposit process. The integration results in a new data management and publication workflow that helps researchers to save time, minimize human errors related to manually handling files, and further promote data deposit together with collaboration across the institution.

Submitted dd Month yyyy

Correspondence should be addressed to Federica Fina, University of St Andrews Library North Street St Andrews KY16 9TR. Email: ff23@st-andrews.ac.uk

The 12th International Digital Curation Conference takes place on 20–23 February 2017 in Edinburgh. URL: http://www.dcc.ac.uk/events/idcc17/

Introduction

Recently, research funders such as the Research Councils UK[1,2] and the European Commission[3] have introduced requirements that expect researchers to properly manage and publicly share their research data. Responsibilities are not just for researchers, but also for institutions which are expected to put in place services to support and facilitate data management and sharing (EPSRC, 2014). Since the introduction of these data policies, institutions have developed local research data management (RDM) services and implemented systems for making research data findable, accessible, interoperable and re-usable, FAIR (Wilkinson, et al., 2016). As a result, some institutions now have RDM services and data repositories in place. Nonetheless, making data open remains a challenge. According to a recent open data survey (Nature Publishing Group, 2016) only a quarter of researchers frequently make their data open (36% when considering the UK alone).

So far the general focus of RDM services and systems has been on addressing the final stages of the research data lifecycle (RDL): archive, share and re-use (Corti, Van den Eynden, Bishop, & Woollard, 2014). However, most of the researchers' activity happens earlier, in the active phase of the lifecycle: collect/create and analyse. Active data are more often managed by using generic note-taking software such as OneNote[4] or Evernote[5], ad-hoc solutions such as specialised electronic laboratory notebooks (ELN), or in-house systems built for specific needs. Full integration of these systems with data repositories is not yet the norm, making the streamlined transition of data from an active to a published and archived status one of the current major gaps of the RDL. Finding a way of filling this gap and providing researchers with processes and workflows across the entire RDL is key to further promote open data practices among researchers.

GitHub and Zenodo offer an example of integration between an active data management system and a data repository[6] (Potter & Smith, n.d.). Researchers that use GitHub to manage their git repositories can login to Zenodo using their GitHub account and submit repositories for publication. Zenodo will then assign a DOI (Digital Object-Identifier) to a specific release. The system can also be set to automatically archive and publish new releases of a specific repository (Potter & Smith, n.d.). The Open Science Framework[7] is an example of a platform that allows researchers to both manage their data during the active phase and publish them at a later stage. Elsevier is working on integrating Hivebench[8], Mendeley Data[9] and Pure[10,11], allowing researchers to link their data across these platforms for better transitioning between active and published data.

[1] Concordat on Open Research Data: http://www.rcuk.ac.uk/media/news/160728/
[2] RCUK Common Principles on Data Policy: http://www.rcuk.ac.uk/research/datapolicy/
[3] H2020 Funding Guide – Data Management: http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[4] OneNote: https://www.onenote.com/
[5] Evernote: https://evernote.com/
[6] Github – Making your code citable: https://guides.github.com/activities/citable-code/
[7] OSF – Open Science Framework: https://osf.io/
[8] Hivebench: https://www.hivebench.com/
[9] Mendely Data: https://data.mendeley.com/
[10] Elsevier – Pure: https://www.elsevier.com/solutions/pure

Similarly to the GitHub-Zenodo integration, we at St Andrews have integrated an in-house developed system with Elsevier's Pure data repository used at our institution (Fina & Proven, In Press). This paper shows how two independent solutions have been integrated with the aim of offering researchers a simple workflow to facilitate and promote the data deposit process.

In the remainder of the paper we discuss the current research data management systems used at the University of St Andrews, present the integration between Pure and NOMAD, and finally, provide content for future work and conclude the paper.

Current RDM systems at St Andrews

NOMAD: NMR Online Management And Datastore

The solution-state NMR (Nuclear Magnetic Resonance) facility in St Andrews is located within the School of Chemistry, but serves the needs of the entire University. The facility is equipped with 6 Bruker NMR instruments and serves 30 research groups (260 users) and about 350 students (i.e. teaching lab sessions' users). On average, about 8,000-10,000 experiments are run every month, corresponding to about 20-30 GB of data.

Managing large quantities of data without a proper content management solution is not only cumbersome and hard, but it also leads to a slow and error-prone research workflow. NOMAD, developed since 2012 as a joint project between the schools of Chemistry and Computer Science at St Andrews, is a web-based research data management system for NMR facilities. Prior to the introduction of NOMAD, experiments were booked on paper, collected data was scattered across different storage units and the task of searching and fetching data represented a significant challenge. NOMAD has become a vital part of our modern open-access 24/7 NMR facility, since it does not only automate the booking process for the acquisition of the data, but it also provides a fast, secure, reliable and searchable data-store, data access protection, and a background job for data replication together with other functionalities that simplify otherwise rather complicated NMR workflows.

Currently, the facility is served by NOMAD v1.2, but work on a new version (v2.0) has started in March 2016 and it is about to progress into beta-testing stage within the first quarter of 2017. Three major upgrades are featured in v2.0: (I) the ability to deploy NOMAD in different environments, such as other universities; (II) enhanced functions to search, manipulate and enrich data from the browsers; and (III) the ability to automatically publish data to a research data system, such as Pure.

Pure: research data repository

The University of St Andrews has used a current research information system (CRIS) since 2006; initially an in-house solution subsequently replaced by Pure in 2010

[11] Elsevier – Putting data management in the hands of researchers with Hivebench acquisition: https://www.elsevier.com/connect/putting-data-management-in-the-hands-of-researchers-with-hivebench-acquisition

(Clements & McCutcheon , 2014). With the introduction of research data policies, in 2015 Pure became the institution's research data catalogue and repository. Since then, workflows for data deposit have been developed with the aim of guiding researchers through the process of archiving and publishing their data (Fina & Proven, In Press).

Even though users are now familiar with Pure and its interface, the research data upload is still a manual process. Often, researchers have to download their digital files from instruments or active data management systems (e.g. NOMAD) before uploading them to Pure via a web browser interface.

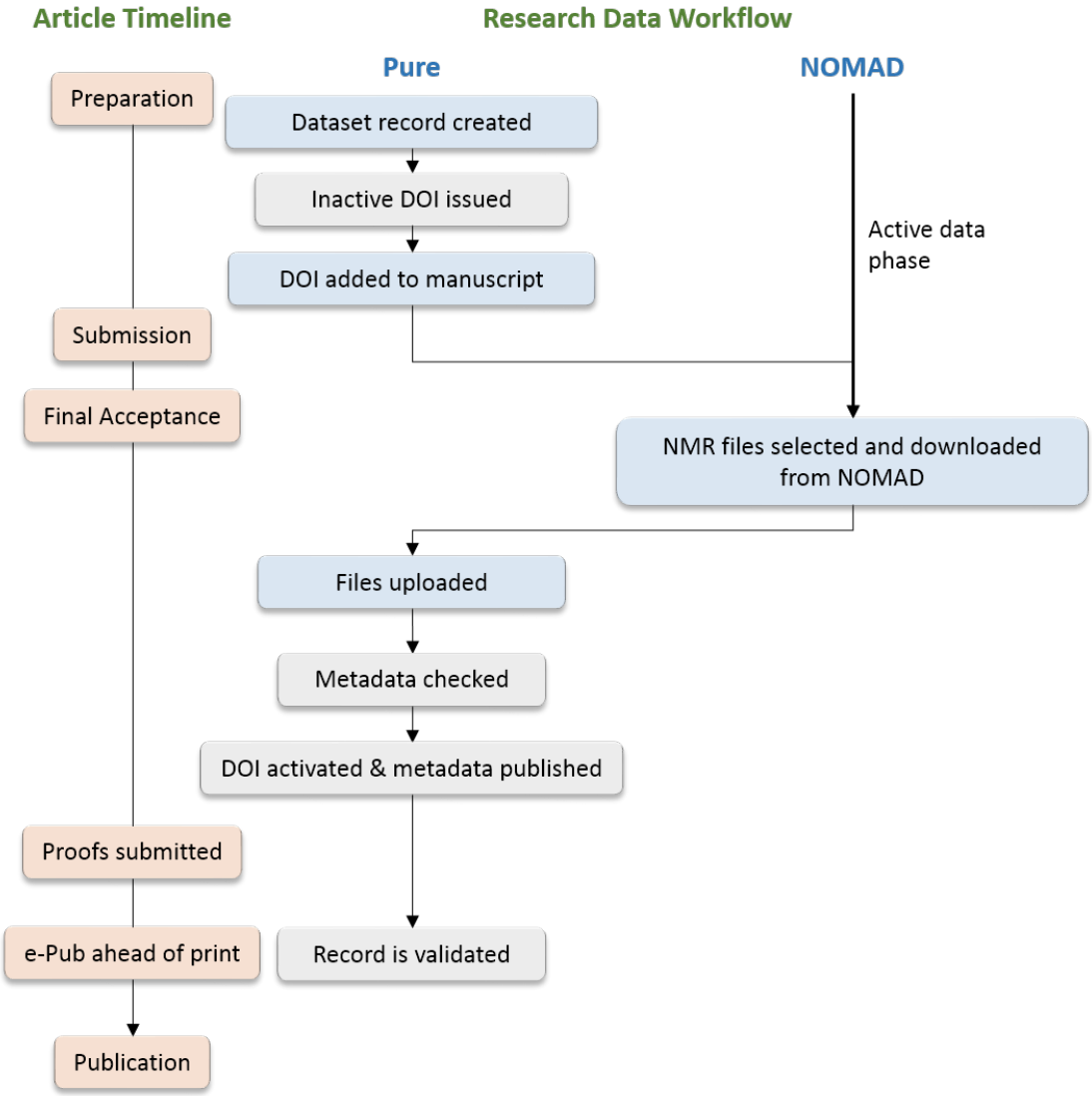

Figure 1 Flowchart representing the current publication workflow for NMR data. Blue boxes = actions performed by the authors in Pure; Grey boxes = actions performed by the Pure system or the RDMT.

The diagram shown in Figure 1 outlines the current data publication workflow for NMR data, also representative of the deposit of other digital files types. The authors create a dataset record in Pure early on in the article writing process; upon record creation the RDMT issue an inactive DOI that the authors can add to their manuscript before submitting it. Once the article is finally accepted after revisions, the authors

download the underpinning NMR files from NOMAD and upload them to Pure with additional data files (if applicable). The RDMT then check that the files and the metadata are suitable and mint the DOI via the DataCite[12] integration (Fina & Proven, In Press). With the minting of the DOI only the metadata are made publicly available on the University's research portal[13]; the data files are released at a later stage, once the article is published.

Pure-NOMAD Integration

Goals and strategy

One of the existing gaps in the research data lifecycle is a smooth transition from the active to the archived and published data phase. The aim of this work is the integration of NOMAD with Pure, with the purpose of streamlining the open data deposit workflow for NOMAD users. The integration must enable researchers to automatically push files from the active data platform to the data repository and thereby removing any manual data files handling.

The current version of Pure[14] does not allow automatic ingest of files from other systems, therefore we adopted a new development strategy that includes an NMR specific open data portal. This approach offers the possibility of tailoring the new portal to the files and discipline type by including NMR spectra visualisation, basic data processing and chemical structure search all with the intent of maximising discoverability and re-usability. Pure still acts as the institutional data catalogue where data records are linked to projects, publications, other datasets and researchers' profiles, DOIs are issued and validation checks performed. To minimise manual actions and automatically manage data files release, NOMAD has been extended to query Pure's API on the status of specific data records, as explained later.

On integrating Pure with NOMAD, we decided that the following basic requirements had to be satisfied:

1. Make the publication process simple and unambiguous for researchers and the RDMT
2. Remove any data download/upload steps in the publication workflow
3. Allow users to specify a licence for the datasets
4. Allow users to embargo published data, if required
5. Allow the RDMT to verify that the data is ready to be published
6. Release only datasets that have been approved for publication
7. Prevent data from being modified after DOI activation and their publication
8. Sign data digitally to ensure data integrity, authentication and non-repudiation

[12] DataCite: https://datacite.org/
[13] Research at St Andrews: https://risweb.st-andrews.ac.uk/portal/
[14] Version 5.7.2-1 as of January 2017

NOMAD and SANO: from data exploration to data publication

NOMAD v1.2, currently deployed in St Andrews, has been designed to allow users to book and program their NMR experiments, automatically save the data into a reliable storage server and allow users to search the results of their experiments. This model is simple and it has proved robust over the years (about 250,000 experiments have been run so far). Until now NOMAD has provided researchers with a dedicated content management system but not a data publication platform. Therefore, an integral component of this development work is an NMR specific open data portal: SANO (St Andrews NMR Open-data). The introduction of SANO is also accompanied by a new data preparation and publication process.

Similarly to the current workflow (Figure 1), researchers initially create a dataset record in Pure, before submitting their manuscript for publication, and the RDMT provide them with a unique inactive DOI to add to the manuscript. Once the article is accepted after revisions, researchers do not need to download the files as in the previous workflow, instead they follow a new publication process as described in the following sections.

Data Preparation Workflow - NOMAD

The most essential steps in NOMAD v.1.2 are data acquisition, data archival and a background backup job as shown in Figure 2 (grey boxes). Once the data is archived, internal users can query the experiments and retrieve the raw data on request.

On designing the Pure-NOMAD integration, it was decided to introduce a data preparation workflow prior to make the data publicly available. The data preparation workflow, introduced in v2.0, consists of metadata enrichment, data packaging and dataset queueing for publication. The metadata enrichment step sees researchers adding descriptive information and attaching chemical structures (useful to visualise and search the data) for the individual and grouped experiments. In the data packaging step, researchers group experiments and sets of experiments together, defining the organization of the dataset in relation to the paper. Finally, authors submit the dataset for publication. On performing this action, authors are asked to choose a licence for the data (e.g. CC BY, CC BY-SA, CC BY-ND, CC BY-NC, *etc.*), specify any embargo period, list the authors of the publication, and add the inactive DOI provided by the RDMT.

Figure 2 Grey boxes = features available since v1.2; Green boxes = features introduced in v2.0; * data is searchable.

The new data preparation workflow provides additional benefits to NOMAD users (researchers internal to the institution) compared to the current content management features. Users are now able to group their data and add discipline specific metadata (e.g. chemical shifts, chemical structures) within the system itself without the need to download the files. The immediate benefits are better organization of the researcher's data and improved data discoverability. Moreover, the research and publication workflows are improved by allowing researchers to collaborate over dataset online,

without the need to download data, exchange emails or using third-party data-sharing software. The goal of adding these new features to NOMAD, therefore, is not only to improve data organization and reduce human error, but also to promote collaboration with other researchers within the institution by facilitating data sharing.

Data Publication Workflow - SANO

After the preparation stage (Figure 3), datasets that are queued for publication are sent to SANO, a lightweight, independent, publicly available, NMR specific open data portal. On receiving a dataset, the data files become immutable and can no longer be modified. SANO then creates an inactive URL for the entry and automatically sends it to the RDMT, who add it to the corresponding Pure record. SANO will then check that the DOI for the dataset is active (i.e. the DOI does not return an HTTP error status) and if so, it automatically publishes the associated metadata. The RDMT perform the usual Pure tasks leading to the publication of the dataset: checking of metadata, activation of the DOI with publication of metadata on the University Research Portal, and, finally, validation of the record. The dataset is finally published once SANO confirms the validation of the record on Pure, by querying its API. In order to do so, SANO needs a unique identifier of the record in Pure. Pure records are identified via UUIDs (Universally Unique IDentifiers), which SANO extracts from the dataset's DOI (doi:institution-prefix/{*Pure UUID*}, e.g. doi:10.17630/*54947df8-0e9e-4471-a2f9-9af509fb5889*). Upon querying Pure, SANO checks the value for the validation status in the resulting xml and if the value is *'validated'*, the dataset is published.

As mentioned, a fundamental and important property of the datasets stored in SANO, compared to NOMAD, is that they are immutable. Immutability is preserved by using a GUID-based content-addressable storage. GUIDs (Globally Unique IDentifiers) are cryptographically hash generated from the metadata and data of a given dataset. Being the GUID uniquely associated with a given dataset, it was decided to map the URL https://www.sano.st-andrews.ac.uk/{*GUID*}[15] to the SANO page of the dataset[16]. The page shows basic useful metadata (e.g. authors' names and ORCIDs, title of paper, dataset description, licence, etc.), discipline specific metadata (e.g. chemical structures for the chemical compounds, chemical shifts, etc.) and a button to download the dataset together with its metadata. The GUIDs allow users to easily check the integrity of the SANO dataset. Often, however, verifying the integrity of data is not enough. Therefore, each entry in SANO is also digitally signed, thus providing integrity, authentication and non-repudiation over the datasets. Integrity ensures that the data has not been modified, authentication allows to determine whether the originator of the data is really who it claims to be, and non-repudiation allows the recipient of the data to send the data and its digital signature to a third-party, with the third-party being confident that the data originated from the initial originator.

[15] Example of SANO URL using sha-256: https://www.sano.st-andrews.ac.uk/*9f86d081884c7d659a2feaa0c55ad015a3bf4f1b2b0b822cd15d6c15b0f00a08*

[16] Raw data will be accessible at https://www.sano.st-andrews.ac.uk/data/{*GUID*}

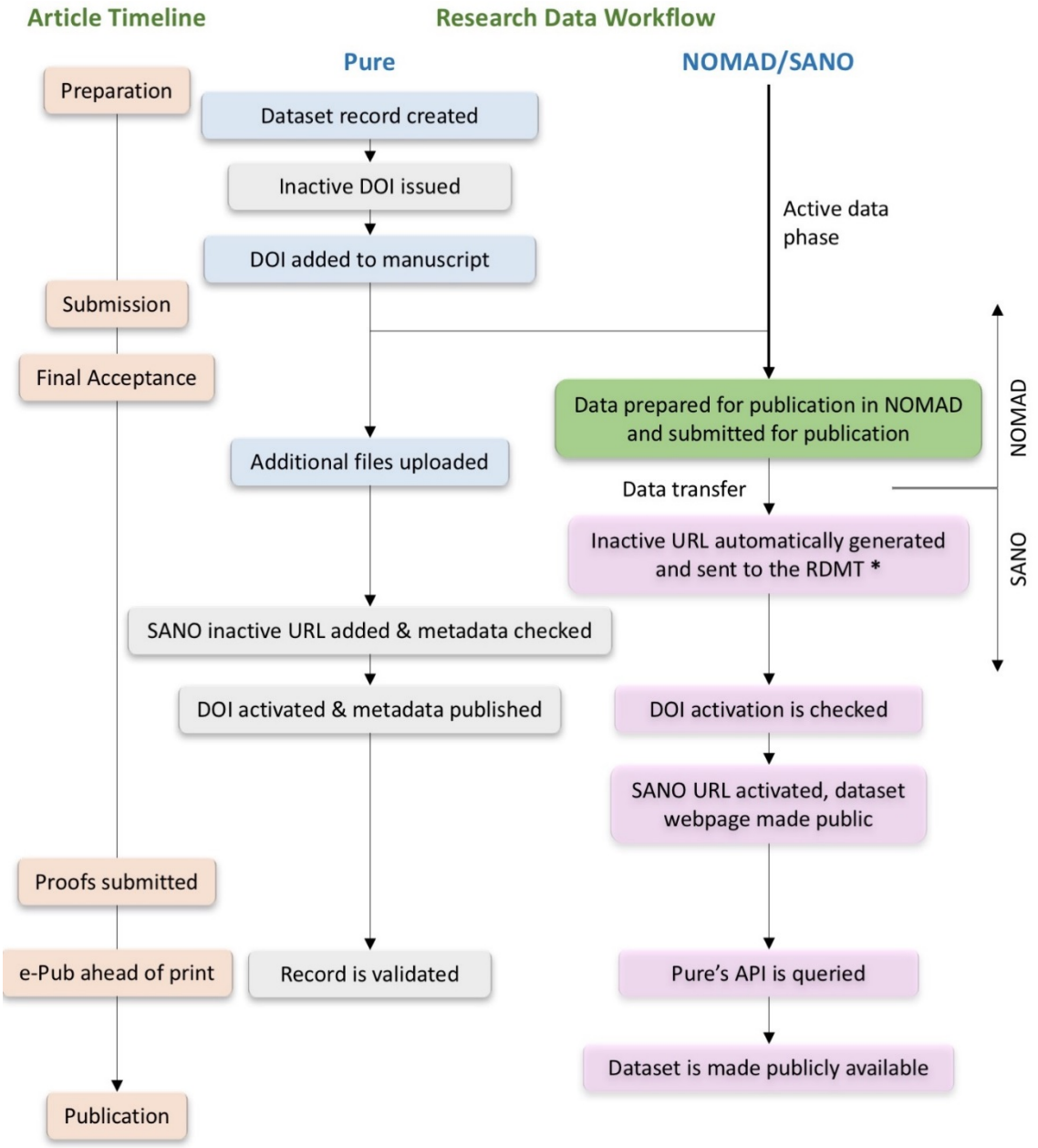

Figure 3 Flowchart representing the new publication workflow for the NMR data, through the NOMAD and SANO systems. Blue boxes = actions performed by the authors in Pure; Grey boxes = actions performed by the Pure system or the RDMT; Green boxes = actions performed in NOMAD; Pink boxes = actions performed by SANO; * data becomes immutable.

Note that detaching SANO from NOMAD results in a simpler design of the systems and helps avoid any interference between the two systems. SANO is currently under active development and it will be released to the community within the first quarters of 2017.

Conclusion and Future Work

In this paper, we have briefly outlined the challenges in publishing research data within research-intensive institutions and provided a general overview of existing solutions to the problem. We then presented the integration between an active data management system and a CRIS system describing the working solution used at the University of St Andrews, where the Pure and the NOMAD systems were integrated seamlessly. We have shown that two existing and already extensively used systems could be integrated together by adding a few additional steps (for the systems) prior to the publication of the data and by automating the data transfer and management steps. NOMAD and SANO offer researchers the possibility of creating and publishing data by using the same platform and therefore reducing file handling. The SANO portal will offer the end-users discipline specific features, such as chemical structure search and visualisation of NMR spectra, while the Pure Portal continues to present data in a rich context providing information about related publications, people, datasets, projects, equipment, activities and impact.

By removing the need for researchers to manually download files from one system to manually upload them into another, the new workflow saves them time and reduces the risks associated with file handling. We, therefore, believe that this new approach will promote data deposit and publication among researchers. In spite of the domain-specific nature of the integration presented in this work, we have reasons to believe that the proposed workflow can be easily generalised and used in another contexts too.

Our upcoming objective is the start of a first beta-testing phase of NOMAD v2.0 and SANO, followed by a school-wide launch and further to other universities. As part of our future goals, SANO will be extended to support advanced metadata search (e.g. by drawing chemical structures) and metadata harvested through Pure's API (e.g. authors, projects, related content). We will also consider integrating SANO with the preservation service that will be developed as a result of the Jisc Research Data Shared Service (RDSS) project[17]. Moreover, the chemical research community currently lacks an NMR specific global repository, similarly to the CCDC portal[18] for Crystallographic data. We think that the SANO portal could be a first step toward building a global NMR database. As part of this, SANO should collect data from multiple NOMAD instances and expose a RESTful API to allow other services (e.g. Jisc RDSS) to work together.

Acknowledgements

The authors acknowledge the work of Juan Karsten for the partial development of the Pure-NOMAD integration. This work has been supported by the EPSRC-Strategic Partners Project (2012, grant number EP/J501542/1) and the Impact Acceleration Account (2016, grant number EP/K503940/1).

[17] JISC Research Data Shared Service: https://www.jisc.ac.uk/rd/projects/research-data-shared-service
[18] The Cambridge Crystallographics Data Centre (CCDC): https://www.ccdc.cam.ac.uk/

References

[journal article] Clements, A., & McCutcheon , V. (2014). Research Data Meets Research Information Management: Two Case Studies Using (a) Pure CERIF-CRIS and (b) EPrints Repository Platform with CERIF Extensions. *Procedia Computer Science*, 199-206. doi:10.1016/j.procs.2014.06.033

[book] Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data.* SAGE.

[report] EPSRC. (2014). Clarification and guidance on the interpretations of these expectations. Retrieved from https://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/

[journal article] Fina, F., & Proven, J. (In Press). Using a CRIS to support communication of research: mapping the publication cycle to deposit workflows for data and publications. *Procedia Computer Science*. Retrieved from http://hdl.handle.net/11366/491

[report] Nature Publishing Group. (2016). Open Data Survey. Figshare. Retrieved from https://figshare.com/articles/Open_Data_Survey/4010541

[web site] Potter, M., & Smith, T. (n.d.). *Making code citable with Zenodo and GitHub*. Retrieved from Software Sustainability Institute: https://www.software.ac.uk/blog/2016-09-26-making-code-citable-zenodo-and-github

[journal article] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*. doi:10.1038/sdata.2016.18