REPORT

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Metadata and research data identification

Author(s):	Thomas Parsons
Audience:	Project stakeholders & JISC MRD
Published:	14/06/2013

Contents

1. Introduction

The ADMIRe project started in October 2011 and aimed to complete research data management pilots in all five Faculties at The University of Nottingham (Nottingham).

The Faculty of Sciences were approached and a number of potential pilots identified. During the summer of 2012 a news release[1] was noticed that announced the conclusion of innovative research to discover a new class of polymers that were resistant to bacterial attachment. These new materials could lead to a significant reduction in hospital infections and medical device failures. Most importantly from an ADMIRe perspective, was that the research was funded by the Wellcome Trust who have an established research data policy[2] and require the data underpinning a publication to be published.

ADMIRe approached the research group and discussed whether they would like to take part in a data repository pilot. With the aim being to ascertain their needs regarding metadata and to eventually share their research data publically.

The following report outlines the findings of the pilot.

2. Nottingham research metadata

The ADMIRe project realised early on that the existing Libraries and Research and Learning Resources (LRLR) schemas would not be suitable for cataloguing datasets. Numerous discussions took place and the work of other institutions was reviewed – particularly the work by JISC MRD Open Exeter project and DataCite[3]. The full metadata schema created by the project is available on the ADMIRe project blog and it is beyond the scope of this report to detail how it was created. However, for the purposes of this work it was important to verify that the fields identified were:

1. Fit for purpose
2. Could be easily understood by researchers
3. Provided the right level of detail

The act of assigning metadata to a dataset would also highlight any issues regarding licencing or identify areas that could be automated or were repetitive.

[1] http://www.nottingham.ac.uk/news/pressreleases/2012/august/new-bacteria-resistant-materials.aspx
[2] http://www.dcc.ac.uk/resources/policy-and-legal/research-funding-policies/wellcome-trust
[3] http://www.datacite.org/

Table 1 illustrates the metadata captured about one of 32 relevant data sets generated by the project:

Institution	University of Nottingham
School(s)	Pharmacy
Research project Title	High throughput micro arrays for discovery of polymers resistant to bacterial colonisation
Alternative titles	Polymers with hydro-responsive topography identified using high throughput AFM of an acrylate microarray
Subject	AFM of polymer microarrays
Primary Investigator (PI)	Morgan Alexander
Lead researcher	Andrew Hook
Other researchers	Jing Yang, Xinyong Chen, Clive Roberts, Ying Mei, Daniel Anderson, Robert Langer, Martyn Davies
Abstract	Atomic force microscopy has been applied to an acrylate polymer microarray to achieve a full topographic characterisation. This process discovered a small number of hydro-responsive materials created from monomers with disparate hydrophilicities that show reversibility between pitted and protruding nanoscale topographies.
Research start date	30/03/2009
Project ID	TBC
Acknowledgements	N/A
Funding body	Wellcome Trust
Funding body reference	85245
Term of funding	4 years
Last day of project	30/09/2012
Date submitted	07/06/2011
Security Rating	Low
Access rights	Public
Date available	TBC
Licence	TBC
Data collection start date	TBC
Last date modified	TBC

Data retention period	None
Audience for reuse	Public
Language	English
Spatial	N/A
Temporal	N/A
Output References	N/A
Data Sources	TBC
Research methods	AFM
Software requirements	Excel
Limitations	TBC
Raw Data Location	TBC
File size	26 kb
File format(s)	.xls
Related datasets	TBC

Table 1: Wellcome Trust metadata

On discussions with the researchers it became clear that many of the suggested fields required further thought and clarification (these are marked by TBC). Of these, areas such as: access rights, licencing and security rating were deemed to require support – after all, the results of the Nottingham RDM survey show that the majority of researchers have never shared a dataset before, let alone considered the type of licence that should accompany it and who may access it.

Although this group of researchers were content for commercial use of the data, interviews with other academics has shown that this is an issue for some – hence there was a need to clarify the meaning of what "public access" really is i.e. access is granted to anyone, regardless of whether they are a researcher or commercial organisation. Similarly the terms of spatial and temporal were not deemed relevant for this type of scientific data.

After a series of discussions it emerged that the metadata schema was missing a crucial element – namely a record of what publication the data related to. This oversight was corrected and the schema decomposed to a core set of elements that were mapped to the Dublin Core as shown in Table 2:

Creator	Andrew L. Hook
Title	High throughput micro arrays for discovery of polymers resistant to bacterial colonisation
Publisher	The University of Nottingham
Publication Year	2013
Identifier	TBC – expected to be a DOI via DataCite
Subject	AFM of polymer microarrays
Research grant code	85245
Location of the raw data	TBC – expected to be a URI or file store link

Table 2: Core metadata schema

Although scarce, the metadata was deemed to be sufficient and was based upon the mandatory metadata of the DataCite schema. Importantly, it provides a link to the paper and the research grant code, which provides the option to automate:

1. Retrieval of project details based upon grant code number from Agresso (Nottingham research finance system) i.e. funding body, PI etc
2. Retrieval of paper details via the DOI
3. Reporting that links grants to research outputs (data sets and papers)

It was then envisaged that the remainder of the metadata fields would be optional.

3. Further work

As the ADMIRe project continued, it became apparent that the technical infrastructure to store and catalogue data would not be available during the project life. Hence there are a number of outstanding actions related to this pilot:

1. Clarify where the data will be stored – the Wellcome Trust project generated approximately 250MB of data related to their publications.
2. Clarify that any data that receives a DOI is stored securely and is archived long-term.
3. Develop licencing models based upon the funding body and type of data generated.
4. Develop subject specific schemas.

5. Integrate the DataCite API[4] and links to the Nottingham Agresso system into any future development work.
6. Upload and describe datasets arising from this project as a test into any new system.

[4] https://mds.datacite.org/static/apidoc