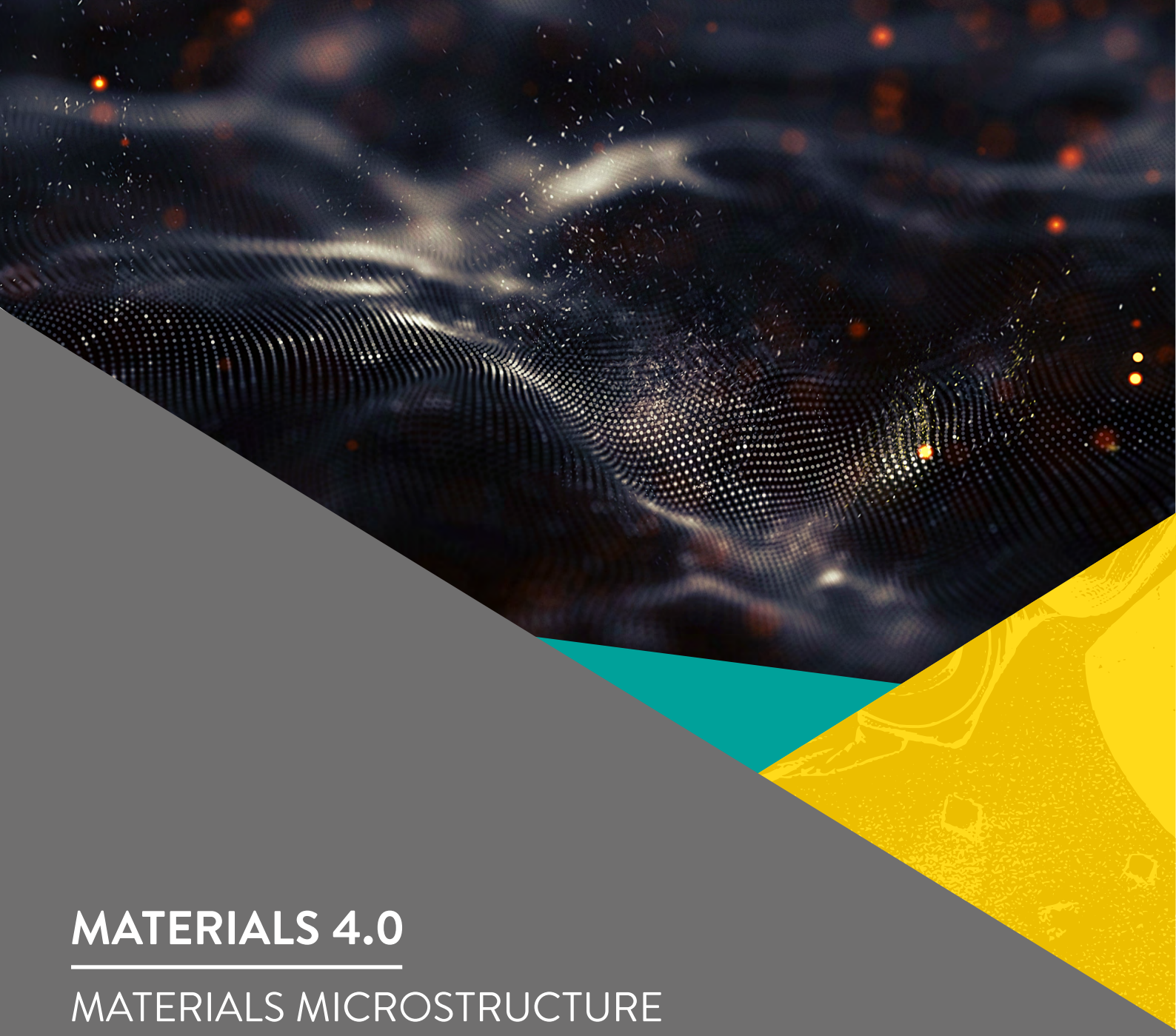

MATERIALS 4.0

MATERIALS MICROSTRUCTURE IMAGE AND DATA REPOSITORY

This report is commissioned by the Henry Royce Institute for advanced materials as part of its role around convening and supporting the UK advanced materials community to help promote and develop new research activity.

The overriding objective is to bring together the advanced materials community to discuss, analyse and assimilate opportunities for emerging materials research for economic and societal benefit. Such research is ultimately linked to both national and global drivers, namely Transition to Zero Carbon, Sustainable Manufacture, Digital & Communications, Circular Economy as well as Health & Wellbeing.

HENRY ROYCE INSTITUTE

Impact Data Metrics

UKRI Engineering and Physical Sciences Research Council

Materials Microstructure Image and Data Repository

Commissioned by the Henry Royce Institute for advanced materials research and innovation.

ABOUT ROAD MAPPING AND LANDSCAPING

This report is commissioned by the Henry Royce Institute for advanced materials research and innovation. It has been published as part of its function of convening and supporting the UK advanced materials community to promote and develop new research activity.

The overriding objective is to bring together the advanced materials community to discuss, analyse and assimilate opportunities for emerging materials research for economic and societal benefit. Such research is ultimately linked to both national and global drivers, namely the transition to zero carbon, sustainable manufacturing, digital and communications, the circular economy, and health and wellbeing.

FOREWORD

Materials 4.0 aims to radically change the rate and responsiveness of materials innovation, increasing the impact it has on society and the economy.

The materials and manufacturing sector forms 15% of UK GDP and has a key role to play in achieving technological and societal goals such as the transition to net zero carbon. Such advances will require the approval and acceptance of materials that are either at the early stage of development or yet to be discovered.

One of the primary challenges to the rate of development for new materials is the poor levels of data sharing between parties. Current working methods result in delays and failures due to a reoccurring need for individual organisations to generate experimental data that has already been generated by others, but is unavailable or difficult to access.

The aim of this paper is to guide the thought processes of the materials community on methods for addressing this barrier. This will be achieved by providing expert opinion on the existing systems available that encourage collaboration and avoid needless, resource consuming, duplication.

Professor Iain Todd, Project Champion and Scientific Lead for Materials 4.0 roadmap, Henry Royce Institute

EXECUTIVE SUMMARY

We have undertaken a review of a number of materials data repositories to support The Royce in its work to develop a resource to accelerate materials science research.

In addition to this, we have taken part in discussions with users representing data users, experimentalists and data modellers/simulation experts.

A consensus received from the groups was that having a place to be able to access quality materials science data was both important and desirable, offering efficiencies and reducing repetition.

Whilst the technology issues are being addressed by various efforts across the globe, the key to success will be user buy-in, with a deep understanding of how to attract users to a new service to both contribute and use being a key requirement.

Having reviewed various repository projects available (both public and private), we have suggest three options:

- Build a indexing/aggregator service - Option 1
- Work with a third party to add to their repository - Option 2
- Build an new repository - Option 3

Option 1 is to build an indexing service for the existing user community. This has the advantage of being able to expand upon the learning of others in this space (see NIST's Materials Data Facility and NOMAD), control the aggregation of data through the index (enabling content to be identified and access provided), and provide The Royce with control as to how such a service is maintained and developed over time. An Indexing service should also reduce the barrier to use through not requiring users to commit data to multiple repositories. It will, like Option 3, require access to third party repositories to be agreed. Over time, and should it become desirable, The Royce may opt to add storage to such a service to ensure that indexed data is made available should third-party services cease to operate. This will bring it closer to becoming a full repository service in its own right.

Option 2 is to work with a third party, which is viable if The Royce is also able to fall under the umbrella of the organisation they work with. This will commit The Royce to an existing set of rules and policies, which could be an advantage in the early days of such a project, will leverage existing expertise and will bring with it access to an existing pool of structured data. The downside of this scenario is that it is less likely that The Royce would be able to exert influence over the future development of such a system and should the third party project cease to be funded, The Royce may find access to the services cut. Finally, coordinating the collection of data to join in the project with content available, will require The Royce to identify a significant source of well-structured data as its contribution.

Option 3 is the most expansive and would require The Royce to develop a full infrastructure and service offering for the materials science community. Whilst this is achievable, there are two main points to consider; (1) Would users use it, and (2) would The Royce be able to guarantee its operation over time? Finally, if

constructed, in order to be of use to researchers, content will need to be available. This will require indexers to be built to access data from other services and repositories.

Our recommendation would be for The Royce to embark upon Option 1 (the indexing service) as it has the potential to develop links in to other repositories and file stores, control its own services aligned with its own policies, and initially reduce the operational and cost overhead of operating a large-scale storage service to curate and preserve data. This option also leaves open the possibility of moving towards becoming a full repository.

Irrespective of which option The Royce choses, there is a need to form a project team to focus on the delivery of such a project, which should be guided by a Panel with both users and contributors to such a project, and report to the The Royce Senior Team and Board.

OBJECTIVE

Impact Data Metrics Ltd ("IDM") were retained by The Royce Institute ("The Royce") to develop an initial view of the requirements for a Materials Research Data Repository ("Repository").

The Royce Institute ("The Royce") is a multi-institutional organisation based at the University of Manchester, covering a wide range research teams and programmes at The National Nuclear Laboratory, Imperial College London, The UK Atomic Energy Authority, and the Universities of Manchester, Sheffield, Oxford, Liverpool Leeds and Cambridge. The Institute partners combine to create a group of more than 900 academics in various fields of materials research and development.

In terms of applications, the research interests of these groups have the potential to impact many industrial sectors across society, ranging from healthcare to energy to manufacturing, and more besides. Cutting across all of this is a common underlying requirement to characterise materials.

Materials characterisation is critical in the development and assessment for their use in new industrial processes. Capturing the conditions of experiments, analyses, manufacture and testing of materials is a key requirement in determining whether or not a material is suitable for particular applications.

Representatives of The Royce and IDM. IDM have had initial discussions about the need for a research data repository to support the wider efforts of The Royce in sharing information to accelerate the development, characterisation and understanding of new materials.

IDM is a Data Analytics company with deep expertise in the assembly, management and analysis of complex datasets. We have built a reputation for working with difficult datasets, using our proprietary AI techniques to gather, correct, and organise data to enable analytics to be undertaken generating meaningful insights. In addition to deploying these capabilities in the assembly and use of our own datasets, IDM provides data repository solutions that enable our clients to take control of their data and leverage the knowledge and insights contained therein.

THE BRIEF

The Royce's interest in developing a system to support discovery of and access to Materials Research Data ('Service'). Initially, it is expected that The Royce will not store images and related material processing data in a datastore, but that it will provide an overarching solution to enable connection to multiple repositories with a clearly defined set of policies and guidelines for its use. Whilst the features required for the Service will be researched in this work, a resultant suggested solution will also ensure that a dedicated repository (cache of data) could be implemented in the future. Given that there will be a dependence upon third party repositories as sources of data, a functional requirement will be to assess and maintain live links and hence there will be a need for The Royce to maintain an index of locations of image and processing data, with attendant search features to enable users to access and find data. An ability to access the 'found' data will be required to enable further research and experimentation to be carried out (e.g. using AI tools for feature identification from image data). In this respect, there will be a long-term requirement for digital curation/librarian expertise to maintain currency of sources and link status and liaise with the content owners.

In summary, this phase of the work is to undertake research to assess current activities in this space to take account of both access to existing materials research data, contribution of new data, together with policies and guidelines to support users.

The required outputs of this work are:

- Requirements capture upon which to base workflows and information content (microstructure, processing and properties data, experimental methods and analytical instrumentation information). This will require a working group(s) to be formed in order to better understand requirements based upon their experiences and needs. Understanding search approaches used by researchers will be required.
- A review of existing materials data repositories for their scope, supported user interactions, and core technologies deployed. This is to determine examples of existing practice and any standards that may be applied to the Service based upon the requirements expressed.
- Development of policies and guidelines that will need to be considered (with examples from other services) for use of the planned Service.
- A roadmap for a project to develop the Service in a step-wise and considered approach.
- This research will be undertaken with a core tenet that whatever is developed, there needs to be a straightforward method for researchers to be able to search and access relevant data, so as to be able to undertake additional experiments. In addition to this, user contribution to the Service (linking to their own data) must be supported. Policies and Guidelines for use and contribution will be developed as a suggested framework for review by The Royce.

APPROACH

The project has a limited timeframe in which to complete. In order to facilitate this, three groups were formed that represented the the key pillars of the research community, which includes:

- Experimental Researchers
- Modelling and Simulation Researchers
- Data Users and Managers/Curators

Members of each group are from the Universities of Manchester, Oxford, Cambridge, Sheffield, Bristol, Birmingham, and Imperial College and The Science and Technologies Facilities Council (STFC).

For each of these groups, an online meeting was scheduled to discuss their current work practices/workflows, their needs, both in terms of data collections and storage, as well as use and re-use of data. In addition, there was a discussion on how each group would like to be able to use the data collected.

In parallel with this, IDM undertook research in to existing services in use in other research groups.

Synthesising these strands of work together, a solution and roadmap for developing, testing and roll-out of a service was to be suggested to The Royce, together with any attendant software and services that would be required.

The findings and conclusions of this research are presented in this report.

GROUP INTERVIEWS

The three groups represented Experimentalists, Simulation/Modellers, and Users. A summary of the key points raised is provided below.

A series of meetings were held to prompt a discussion across three groups; Experimentalists, Simulation/ Modellers, and Users. The general areas of discussion centred on:

- Data Generation
- Data Processing
- Data Storage
- Finding Data
- Finding the Right Data
- Data Sharing

Using these key points as a framework for discussion ensured that we are able to garner some consensus of views across the different groups, and allow a free-flowing discussion in to other areas of activity.

All the groups noted that a typical starting point for a research question began with a literature/database search. Whilst this is time-consuming, it has in some instances been automated using Python scripts. Some comments were made about research students not being able to find data and having to generate it *de novo*, despite Principle Investigators (PIs) knowing it has already been done by other groups but not yet published to repositories. This was echoed by the Users group, who stated that data sparsity and volume is often an issue, particularly for modelling purposes. Of course, such data sparsity may be due to the dispersed and fragmented nature of existing repositories themselves, and knowing where the 'right' kind and quality of data is, is hard to establish. This highlights a number of key points:

- Data may already exist but it is simply not accessible
- Search is complex and requires a researcher to look across many different sites
- Quality of data is potentially variable or even unknowable

In relation to workflows, it was indicated that whilst they do exist, and share some common features across experiments, there are a multitude of methods for collecting and processing experimental data alone. Whether that is due to different instruments (even within and between manufacturers) or the different processing of a material prior to analysis, standardisation is problematic. In short, capturing all of the potential use cases becomes a Sisyphean task. This was exemplified by the Simulations and Modelling group and experience garnered within the LigthForm Project. Despite the latter being focussed across a limited number of research

teams, the project was not without its problems. As part of the LightForm project, there was a programme to collect metadata for a wide variety of instruments that experimental teams were using. This proved to be an intractable problem to solve both in terms of generating all of the required metadata schema and having users input the data, with the former referred to as an 'explosion' of schema. This led to discussions around binning metadata in to classes that could be collected automatically versus those that needed human input. It has been tried within LightForm and was indicated as being easier for physical experiments than for simulations, but having a non-materials scientists, or even a materials scientist who had no experience of the analytical method used, hampered the decision-making of what metadata was important or not. That said, it was also acknowledged that the work in the LightForm project represented progress in this space. It was also noted, that with simulation/modelling experiments, there were a some Python tools that supported researchers in metadata cataloguing for experiments, but once that experiment deviated from set parameters, the utility of tools reduced and they were abandoned. At the University of Sheffield, similar progress was being made in relation to the automated collection of machine-derived data, which has created a standardised approach that to some extent is helping to support materials discovery by ensuring data is made available to the AI/ML researchers more quickly and with a more consistent set of metadata. In both cases described by the groups, data storage was also linked to the project. This is a clear advantage of such focus in these projects, but translation up to a larger grouping such as that represented by The Royce, or even a wider, future programme run by The Royce on behalf on the UK materials research community, may prove difficult as the numbers of users increases and cultural practices of different groups comes in to play.

Finally, in working through this part of the discussion, the question as to what data is required in a repository was focused towards Open Science, not reproducibility. The logic being that a research publication should contain sufficient information to recreate the experiment, and therefore only the metadata is critical post-publication, not the raw data *per se*. Additionally, the quality of metadata is important and feeds directly in to the provenance question. This is a central feature of a long-term digital repository curation function, upon which there will need to a focus.

The key points from this discussion on metadata were:

- Metadata is important
- Quality of metadata is important, and therefore curation.
- Metadata has many standard features, but due to the inherent variety of analytical techniques and machines, and well as the novel nature of research, automated collection and classification of data remains a problem.

With respect to user buy-in to using a repository, this was cited as being a significant barrier for any such project. There was discussion around such barriers, being:

- A lack of appreciation of the value of the use of data beyond a publication or its original purpose
- Sharing data across a community
- Concerns around security and control of the use of the data

- Making a decision as to the point at which data is shared

These points clearly underline a general problem, which is incentivising researchers to use a repository service. As with any human-system interface, it is better to show benefits (carrots), and only use 'sticks' if they are necessary. One potential benefit indicated was being able to search for and find data. As discussed above, searching for data is a laborious process, and any improvements that could reduce the time and provide access to quality and accurate data would be a boost. Other points raised included ideas around insisting that there is a data management plan prior to equipment use. It was suggested that The Royce could be part of this across its projects and partners, albeit that this in itself may take time to introduce and require careful consideration in its implementation. Another suggestion was to introduce training and education programmes for data management for research students and postdocs, as well as the scientists that operate instruments. This is very much a change management process and would take time to develop and bed-in.

One thing is clear, from our own experience as developers and users of repositories — acquiring community buy-in is key. This requires both a cultural shift and training. A critical hurdle for this is time as users may already be adding data to institutional repositories, and may not want to spend the time adding it to a second. Now, whilst some funders are considering holding back part of the research grant as an incentive to have recipients actively add data to a repository (the 'stick' approach), that may not be available to The Royce, which is ultimately seeking to create a resource for the benefit of the community. Also, this type of payment retention approach will only serve to inflate grant award values over time as organisations will simply build project budgets and discount the retained amount to offset risk, as happens in the construction sector.

The key points on this discussion around users were:

- User buy-in is key
- Barriers to entry are high and would need to be lowered
- Benefits of contributing to a repository were acknowledged
- Culture change would require continued and persistent training to be introduced
- Users would need to see and appreciate the benefits
- Security is a concern

With respect to data size, discussion revealed that data files from experiments could be in the 100GBs to 10TBs. At this volume, moving data around becomes an issue as well as tracking data across its 'lifetime' of use. Often, the process of transfer came down to simply dumping machine data on to physical storage media and physically moving the data to another machine for processing and analysis. From there, it may stay on the external device or be moved to a cloud storage service (DropBox, Google Drive, AWS etc), or locally provided file store or repositories. This can make finding and re-using such data problematic. For example, given the transient nature of postgraduates and postdoctoral staff, many of whom may leave the field entirely, tracking what was done in an experiment after they have left becomes very difficult. An added 'headache' taking this

ad hoc approach to data management is that the individual PI of the group has to determine a data retention policy, which often as not is driven by whether a piece of work has been published. Conversely, there may be other cases where data retention is mandated externally. In such cases, the task of reviewing 'old' data is time-consuming and either neglected until it must be addressed because, for instance, the storage space is needed for new data, or it is deleted based on arbitrary rules, such as 'will I need it again?'. Whilst this may not appear to be a satisfactory data management/retention policy, it is clear that it is an expedient method of working due to the absence of another, easier to use and manage solution. One interviewee suggested that the storage and retention of Climate Data and Models[1] may be a useful comparison.

This indicates a number of key points:

- There is no fixed methodology/policy for the storage of data files, nor the metadata
- Data can exist in multiple places, even within a research group
- Tracking the data over its lifetime is problematic

Summary of Key Features

Feature	Key Points
Data Generation & Processing	• There exists a variety of instruments, manufacturers with proprietary formats for storing analysis data. • Attempts have been made to create data collection forms and schema, but they are numerous and diverse • There are consistent metadata across many experiments, but likewise due the the very nature of research and experimental design, there are also numerous non-standard features. This makes it difficult to standardise all process for a data/metadata collection workflow that is generically applicable to materials science researchers. • Determining the importance and quality of metadata from instruments and experiments is often not a job that can be handed on by a researcher to a third person. • There are a general lack of people dedicated to managing the operation of machines that have longevity and could play a role in supporting Data Management policies.
Data Storage	• Data files can be up to 10s of TB depending on the instrument and experiment. • Big problem is data transfer from instruments to analysis workstations, even with fast networks. • Data is stored in a variety of locations and media types (USB sticks, external hard dives, users machines, cloud services and tape (HPC service operators).

[1] https://ncas.ac.uk/our-services/computer-modelling-and-data/centre-for-environmental-data-analysis/

Feature	Key Points
	• There is often no clear workflow in place outside a research team to standardise data storage and cataloguing.
Data Management	• Data retention policies are often *ad hoc* and depend upon individuals to make decisions as to how long, and for what purposes, data files are retained. There are some projects on-going to address these issues, but from within a project group. These are finding variable success. • Data curation is a full-time job, and not one that will likely yield more recognised research outputs.
Search & Quality Assurance	• Lack of a unified method of search for data recognised by the fact that researchers often begin by searching through existing research papers, or begin with a Google search. • Data may already exist, but location, accessibility, quality and provenance may be difficult to establish and thus, unknowable, which may affect results ranking in returned searches. • Accuracy of search. Need to be confident that what you return as a hit in a search is what you are looking for. • Classification of data is a problem. • Quality of datasets from other sources often needs to be assessed.
Sharing & Security	• Notwithstanding the need to embargo data prior to publication or for contractual reasons, sharing was accepted as a good way to support the wider community. • Generating citations and enabling research by leveraging existing data is a benefit. • Users must be able to control access to their data. • Mixed views were received in relation to sharing, IP and competitors. • Private sector partners often not willing to share data. • Third party storage services (e.g. AWS) were used in some cases, but in some cases, these were not considered private.
Users and Culture	• There must be a low barrier to using a system • Culturally, users are focussed on running experiments and analysing the data, not keeping detailed records in a data management system. • For users, the key thing is the publication output, which is a highly competitive activity, and not necessarily aligned with good data management practices. • Mandatory conditions could be applied, but there was a difference of opinion on this. • Training would be needed to change culture and mindset.

Overall, there was a view that there is a need for a searchable repository, with low barriers to use, secure, with control of release of data. The discussions very much underlined two main points; (1) something of this nature is needed to speed up research, (2) cultural issues are very much the barrier to adoption.

RESEARCH - EXISTING SERVICES

Materials research data is both expensive to produce and invaluable to the research community. As with may types of research endeavour, it begins with the formulation of a hypothesis to be tested, experimental design to test the hypothesis, experimentation, data collection, analysis and reporting. This takes time, both researchers' time and machine time, as well as materials expenses. The outputs of these experiments are rich and valuable datasets that include:

- Experimental conditions and protocols
- Experimental information about the material created generated by analytical instruments
- Meta-data relating image, machine parameters and other data-related information (owner, creator, timestamps, size, location, etc)

Major drivers in creating repositories include:

- Accessibility of data
- Longevity
- Re-use of data for other experiments (e.g. for modelling)
- Reduce the need to repeat experiments
- Sharing of data
- Fidelity of data
- Provenance tracking
- Building relationships, particularly through shared interests

An initial focus has been suggested for micro-structural data, process data and property data.

A number of materials data repositories are currently available on-line (either free or under license), and in undertaking research on these systems, and in so far as we can ascertain, we will focus upon:

1. Ease of management of services that will be ultimately deployed
2. Flexibility to change and/or introduce new data types and themes
3. Ability to provide a long-term curation services
4. Staffing levels and skillsets to operate such a service
5. Security model
6. Digital curation requirements, with an emphasis on quality control of data entered

Whatever the state of the information available for this research phase, we will be looking to develop these features in the solution for The Royce.

MATERIALS DATABASES REVIEW

In this section, we report the findings of our research on existing Materials databases, which will include databases that also hold micro-structural information. Whilst we are focused on planning a solution for testing by The Royce, we are not necessarily interested (at least at this stage) in the underlying technologies that drive the systems. For example, identifying the specific backend database technology provider is less important that understanding the features and content.

Total Material

Total Material (www.totalmateria.com) is a subscription database of over 450,000 materials (Metals, Polymers, Ceramics and Composites), claiming 3000+ sources for advanced data, 74 Standards Development Organisations, 150,000+ stress-strain curves and 35000+ materials with cyclic properties.

It provides a searchable interface through a Quick or Advanced Search option. Using the Quick Search option, a material designation code can be entered, and further filtering of results can be achieved by selecting a country and a standard. Results are returned in a tabular format with hyperlinks to additional information. The table output generally returns the material code, Standard, Country and Type of material (e.g. Metal, Ceramic etc). Clicking on the desired material from that list opens a page with more specific data entries for the material, which is summarised below.

Subgroup	Properties
The material found by is code and any further filtering by Standard (e.g, EN standards)	Properties, which may include Cross Reference Tables, Composition, Mechanical Properties, Physical Properties, Heat Treatment

By clicking on to a property, the browser forwards to a page with details therein, e.g. Chemical Composition.

The cross-reference option links the material to equivalence data for other standards for that material code. Equivalence include referencing links to identical materials, by composition, and access to a tool that enable comparison between equivalent materials to be undertaken.

A final feature to aid search is called SmartCross2, which allows a user to filter equivalent materials by degrees of matching to chemical composition or mechanical properties of materials. This feature, together with a similarity threshold option, are a proprietary function.

Advanced search options open up the database for search using parameters of interest to aid in finding materials, such as specifying chemical composition or mechanical properties, etc.

They offer the core services for materials in the Total Metals database, to which they have added a number of functional modules ranging from Environment characteristics, Compliance, non-metallic materials, supplier directories and so on. They operate a commercial licensing model.

Deployment is dependent on the customer and can be both on-site using a customers own hardware and network (more appropriate for larger corporate clients) and accessed via a cloud service.

There is no information on the curation policies *per se*, but one might suppose that they will be simply growing the size of their repositories through continuous integration of new records.

MatMatch

MatMatch Like Total Materia, MatMatch (https://matmatch.com) offers access to a database of information relating to materials properties built around a corpus of 31,000 materials. As with Total Materia search is via filtering by material types and properties. It allows comparison of properties of materials through a variety of charts and tables.

GrantaMI (part of Ansys; see https://www.ansys.com/en-gb/products/materials)

Both simple and advanced searching, providing materials. 5 data sources feed in to the system, which include Alloy Finder, Data Sheets and Diagrams, Materials Property Data, Corrosion and Performance Data, and Coatings Data. Searching via similar composition or chemical properties is available, with US and International Standards provided. It draws data together from 21 different sources and provides tools for comparing and matching materials to requirements as well as access to materials data. The company was founded in 1970 as a simulations business. Declare themselves as being particularly strong in the simulations space. They have specific access products for students (EduPack) and digital learning with 1.5M+ students having downloaded their products, with 1500+ Universities using their products. Their products are built to focus upon key technology development areas for industry sectors, e.g. Automotive, Aerospace, Energy and so on. Strong AI/ML component to their overall offering, that generates new tools to leverage the datasets they hold. They are an acquisitive company. Their high-performance compute platform operates with up to 960 cores and access to GPUs (Nvidia)

Whilst these offerings provide an indication of what is available in the commercial space they are very much focussed on supporting a commercial R&D and material sourcing supply chain. As such, while their interfaces for searching are instructive, their utility within the context of research data repository are limited.

ASM International (https://asm.mpds.io/#start)

ASM International is a membership-driven organisation based at Materials Park, Ohio, USA. They own digital libraries of data held within a secure repository which they use as the basis of their own products (e.g. ASM International Materials Platform for Data Science and ASM Alloy Center Database) or license access to to third parties, including Ansys/Granta. They have over 3M records for materials taken from peer-reviewed work. They charge anywhere between $250 and $2,200 for a single seat user license per year, based on which of the 8 datasets a user subscribes. Whilst they operate as a foundation, they are primarily a content publisher. They have been involved in a number of initiatives related to data curation of materials sciences data, and founded the Computational Materials Data (CMD) Network in 2012 to support the collection, management and dissemination of materials data. It was launched in response to the US Materials Genome Initiative. It began this endeavour (with partners) and utilised the Granta software to build its open materials data resource, and ensure that there was sufficient capacity to maintain it as a long-term, curated digital repository. A Structural Materials Data Demonstration Project was launched as a pilot and used DSpace for its repository solution with GrantaMI software solutions. The CMD Network has since evolved and ASM International is a partner in the CHiMaD project (see below).

The Centre for Hierarchical Materials Design Project

The Centre for Hierarchical Materials Design (CHiMaD), a collaboration between Northwestern University, University of Chicago, Argonne National Laboratory, QuesTek Innovations and ASM Materials Education Foundation (part of ASM International) and funded in part by the US National Institute of Standards and Technology. Its mission is:

"Accelerating materials discovery and commercialization by design and development of hierarchical methods and materials and enabling the complete integration of theory, computation and experimentation by building a strong community of current and future researchers."

CHiMaD links to a variety of projects Materials Genome Initiative services and databases beyond the MDF, including a registry of materials resources (Materials Resource Registry) of 35 different types of material databases, different tools for and libraries too support researchers, and assets to support ontologies across a variety of engineering-related fields.

Beyond their own research activities, CHiMaD supports the materials research community by providing access to a number of databases and tools for data mining, including the National Institute of Standards and Technology (NIST) Materials Data Facility (MDF). The MDF is a combination of services; Publish, Discover and Connect. It was established to ensure that data generated through publicly-funded research was made available to encourage its use and re-use of data, and access to a range of discovery tools.

MDF Publish is a decentralised data repository that enables users to publish their content from a variety of storage services (Google, DropBox, Box), repositories (e.g. FigShare, Zenodo) to any Globus (see, www.globus.org)[2] endpoint hosted at the University of Chicago, providing a persistent (DOI) identifier and initiating data curation workflows. They use 4CEED to automatically curate uploaded data whilst running metadata extractors. MDF Discover is a scalable, access-controlled, cloud-hosted search index with tools for advanced search and retrieval of records. Finally, MDF Connect is the service that sits between MDF Publish and MDF Discover. It has 3 primary roles:

1) To connect user queries to the locations of stored data for retrieval
2) Enrichment of data collected from data stores (e.g. Google) using general and materials-specific metadata extraction tools, and
3) Dispatch of data to MDF Publish or to other community data services selected by a user.

As of their 2019 Annual Report, they had over 230 datasets, with over 50 from CHiMaD, covering greater than 350 authors and storing over 45TB of data. At that time, many of the data extractors they had developed were for structure data, they recognised that there is a significant body of work in the literature that researchers often use. They recognised also that extracting data from publications is a challenge, but they have made steps in developing Natural Language Processing tools (SciNER a generalised neural network model for tagging scientific entities in articles) that had a 50% performance improvement on existing state-of-the-art tools of the time.

Access to services is either via programmatic tools (Python) or a more recently developed web interface. To assist with automation, a REST API is available. The search features in the web interface enable full-text queries, partial text matching, and more advanced queries. It is designed using FAIR Data Principles[3] (Table 1).

FINDABLE
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
ACCESSIBLE
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable

[2] Globus is a secure research data management service for moving, sharing and managing non-federated datasets with over 30,000 end points connecting in to various storage location types (tape, HPC, local storage, etc), with a user security model and a developer API toolkit. It was originally developed in 1997 to support data management on grid computing.

[3] Wilkinson, MD et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, **3**, Art. 160018

FINDABLE
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
INTEROPERABLE
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
REUSABLE
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Table 1. FAIR Data Principles.

CHiMaD is now in phase 2 of it programme, which runs to 2023.

NOMAD

Formed from an EU-funded project (grant number 951786) initially running from 2015 to 2018 and now extended to 2023. The consortia is lead by the Max Planck Institute (Fritz Haber Institute; Munich) and 11 other universities, including Cambridge (Csányi Lab) and Warwick (Kermode Lab) in the UK. The overall budget was €4.9M. NOMAD Laboratory[4] is a service that was developed to support the curation of computational materials science data/metadata (10,968,632 entries at 21 March 2021), with a focus upon simulations, and is developing its offer for experimental data (283 entries) and quantum computer data (1 entry). It has a well-developed set of metadata schema. The FAIR data structures employed in the system were developed specifically to established the readiness (quality) of data for AI experiments. NOMAD is free to use, aims to store data for 10 years, and provides a DOI for each data record to support citation. NOMAD Tools are described as:

- NOMAD Repository - climbing over 100million input and output calculations
- NOMAD Archive - a standardised, homogenous representation of data in the repository built for open-access.
- NOMAD Encyclopaedia - for advanced search
- NOAM Artificial Intelligence Toolkit - for enabling new discoveries

[4] https://nomad-lab.eu

For the user, NOMAD has a well-developed interface and a query language provided. It also integrates with data from The Materials Project (Lawrence Livermore Laboratories), AFlow[5] (a consortium of 16 US labs) and Open Quantum Materials Database[6]. With respect to these services, there appears to have been little current activity in terms of publications outputs, with papers listed on their respective websites lacking information from 2019 onwards. This indicates that these projects may well be continuing but only as an internal activity to those partners, or that the initial funding support may be dwindling and maintenance and development is proving problematic. This highlights a key aspect of digital repositories, which is the ability to establish a long-term digital curation service with a model that continues to attract more users and content. Without this, it can become a niche library of data that could be forgotten as new projects are initiated. In the first period of the project (2015-2018) it generated 60 papers from users' data.

NOMAD stores data at the Max Planck Computing & Data Facility and the software is, in essence, and SaaS toolkit. For a user, adding data is restricted to files of under 32GB at a time and only 10 non-published records can be held on the system by a user as any one time. At upload, a user is able to manage the the process of metadata extraction and publish to the repository once they are satisfied. At that point, the user can choose to publish the data and make it available to others. The system also has a complement of Python libraries, schemas and scripts to support various facets of activity either as developers, system administrators or users.

NOMAD also provides an OASIS product, which is a stand-alone version that can be deployed within an organisation and connecting to internal data stores. All the software is open sourced and the code base is maintained. It is available for use at a cost of €2000 per year for an Academic license.

A summary of these repositories is provided in the table below.

Service	Feature	Notes
Total Material	Management	Aggregator of Data. Data hosted securely on top tier provider. Has partnerships with organisations to provide content, including publishers such as Elsevier (Knovel).
	Flexibility	Multi-language support, users restricted to 2000 pages per month.
	Long-term curation services	Product has been evolving since inception in 1999
	Staffing levels and skillsets	Team of 25 skilled, customer-facing staff. Four people are listed in the senior team, and would estimate as many as 15-20 developers/data scientists, and admin/accounts staff of a similar size.
	Security model	ISO 27001 Certified. User registration is required.
	Quality Control	Multiple ISO 9001 Certified processes. Relies on 3rd parties for quality.

[5] http://www.aflow.org

[6] http://oqmd.org

Service	Feature	Notes
	License	Commercial. Proprietary software.
	Terms of Service	https://www.totalmateria.com/page.aspx?ID=TermsOfUse&LN=EN
	Business Model	License fee for access. Publishing model.
MatMatch	Management	Aggregator of Data. Sources include M-Base, MakeItFrom.com, Materialsgate, WIAM, Xincailiao.
	Flexibility	They enable suppliers to list their products on their system together with relevant data.
	Long-term curation services	Established in July 2017.
	Staffing levels and skillsets	With a net cash position of €9.8M year end Feb 2020 and a burn rate of ca. €3.5M we would estimate a staff complement of 25-30 people. They list 8 key people.
	Security model	Uncertain beyond user registration process.
	Quality Control	Uncertain, but would suggest that quality control is with the creators of the content.
	License	Free to users. Suspect the model is a charge to materials suppliers to list their goods, and so this is a cataloguing service that has a place in research, but not a research repository *per se*.
	Terms of Service	https://matmatch.com/static/pdf/Supplier_Terms_and_Conditions_EN.pdf
	Business Model	Research indicates the model is a charge to materials suppliers to list their goods, and so this is a cataloguing service that has a place in research, but is not a research repository *per se*. Publishing Model.
GrantaMi	Management	Well-developed service. Aggregates datasets on to own cloud/HPC platform.
	Flexibility	Draws data from over 20 different sources. Provides access to data and advanced search tools. Data and tools are accessed via their cloud.
	Long-term curation services	Company has been running since 1970, providing the stability of operation required for long-term digital curation.
	Staffing levels and skillsets	Part of Ansys, a $1.5billion company listed on the S&P500 and Nasdaq. Over 4,800 employees across multiple territories, with 150 channel partners.
	Security model	Uncertain beyond user registration process.
	Quality Control	It draws its data from multiple sources. Quality of data is a third party issue.

Service	Feature	Notes
	Terms of Service	https://www.ansys.com/content/dam/legal/wla-august-14-2020.pdf
	Business Model	For profit licensed access model. Publishing Model.
ASM International	Management	Membership-driven organisation managed by a board of trustees
	Flexibility	Data is mostly derived from peer-reviewed articles.
	Long-term curation services	As an organisation, they are over 100 years old
	Staffing levels and skillsets	Will be linked to the publishing activity of the organisation.
	Security model	Uncertain beyond user registration process.
	Quality Control	It draws its data from multiple sources and peer-reviewed publications. Quality of data is a third party issue.
	License	Commercial, but fees linked to membership of ASM International.
	Terms of Service	https://www.asminternational.org/documents/10180/0/ASM+Corporate+Member+Database+License+Agreement+%2807-16-19%29+%281%29.pdf/c0a1933b-2fb3-1540-8abb-633a3d3503f5
	Business Model	Subscription model for each of their datasets. Publishing Model.
NIST MDF	Management	Sponsored by NIST and CHiMaD (University of Chicago-led)
	Flexibility	Service structure appears to provide a great deal of flexibility allowing researchers to add data to the repository from a variety of sources. Computational and experimental data.
	Long-term curation services	The longevity will be entirely dependent on the main sponsors (University of Chicago and NIST), albeit there is current funding and the utility of the service has grown. Earliest published dataset is 2013. Longevity appears stable, but this will be affected by changes in national industrial strategies.
	Staffing levels and skillsets	Multi-organisation, multi-disciplinary teams of developers and researchers working on discrete aspects of the service.
	Security model	Workflows enable users to control when data is published to the repository. It is open access to use. Restricted access to data set by contributor.
	Quality Control	Reliant up data generators.
	License	Free access to and and contribute.

Service	Feature	Notes
	Terms of Service	https://materialsdata.nist.gov/page/tos
	Business Model	Part of the US Materials Genome Initiative.
NOMAD	Management	Consortium of 12 Universities.
	Flexibility	Open access. Only computational data. Plans to develop the experimental data element of the system. Available to users to contribute to the central NOMAD repository at Max Planck, or license the software (OASIS) for use in their own organisation.
	Long-term curation services	Project has been running since 2015. Indicate a 10-year limit to storage of user data, and 3 years on embargo of uploaded data.
	Staffing levels and skillsets	Multi-organisation, multi-disciplinary teams of developers and researchers working on discrete aspects of the service.
	Security model	Workflows provide control by data owners on submission.
	Quality Control	Quality control is in the hands of the data creators.
	License	Free to use central NOMAD service. Separate, local version costs €2,000 per year for an Academic license.
	Terms of Service	https://nomad-lab.eu/services/terms
	Business Model	Uncertain. It is a project funded centrally, and will therefore by subject to funding through Universities and central EU funds.

With respect to policies, the terms of service provide some interesting information. All of the services do not warrant the quality of the content the serve to users, which whist it is sensible to operate in this manner, it means that the issue of the quality of the materials data is unknown. A question that arises from this is whether a service deployed by The Royce should have a ratings system based on a metric that the user community themselves can judge?

Instructions for user contributing to repositories varies based on the operating model employed by each of the systems reviewed. Those of direct interest for this project are for the repositories operated by NIST MDF and NOMAD.

The MDF provides access to a variety of schema and tools to support user submission of data, and access to the MDF itself. These are made available at https://github.com/materials-data-facility. Of particular interest is the MDF Connect code repository section providing example of code used to harvest data from other repositories, parsers etc.

NOMAD provides full documentation at https://nomad-lab.eu/prod/rae/docs/introduction.html. This includes full guidelines on both development of code, their data model and information about data uploads, including

file size limits of 32GB, only 10 non-published uploads per user, a 3 year embargo limit and a 10 year lifetime in the repository. Whilst these appear to be straightforward data management and retention policies, for this project, there needs to be a view taken as to whether or not The Royce will be committing to building and running a full repository service, or acting as an indexer (see Recommendations below).

That notwithstanding, it would be still be prudent to have policies regarding the metadata about datasets, (quality, assurance and completeness), whether the dataset *per se* is held or referenced.

THE (ALMOST) UBIQUITOUS NATURE OF REPOSITORIES IN ACADEMIA

Repositories are, in theory at least, an example of good data management practice to store information in a searchable and accessible manner in to the future. Notwithstanding data retention policies, which will depend on the funding requirement of research councils, etc, legal requirements, and practical matters such as cost of storage, they are used in many of the partners of The Royce.

That said, the repositories in use are varied, both in subject matter stored, and underlying repository systems. For example, at Manchester, they have recently announced the adoption of FigShare for their research data management, whilst the LightForm Project in Manchester has opted to use Zenodo. At Imperial, they use DSpace, linked to a front end from Symplectic Elements, which is also used by Cambridge. Oxford also uses DSpace, as does Manchester (albeit that it is not certain whether FigShare is a replacement). In short, even amongst the partners, there is a broad selection of repositories in use already.

This brings with it a number of observations and questions:

1. Is there a need for yet another repository?
2. Given that researchers may already be using institutional repositories, should we add another one?
3. Do researchers actively adhere to institutional data management policies?
4. Do researchers add their data to their institutional repositories?

This really highlights a fundamental hurdle in developing projects of this nature, which is engagement of users. Users will drive the need for this service, they will be contributors, either directly via their own research, or through existing institutional repositories. They will be users of the information contained within the repository and will, therefore, need to be assured of the quality of the data entries. Our interviews indicated that this was a good thing to be done for Materials Research, but also recognised the difficulties of engaging researchers to contribute and use the repository. Main elements expressed were around:

- Time
- Security
- Control & Ownership
- Value to the researcher

OUTSOURCING OF REPOSITORY, RESEARCH DATABASES, AND META-REPOSITORY PRACTICES

In many fields in Academia the most ubiquitous and long-lived repositories are a collaboration between publishers, libraries, foundations and academics. The involvement of foundations, such as the Andrew. W. Mellon foundation, provides seed-funding for the establishment of new collections, whilst the collection management, technology expertise and dissemination comes from organisations such as JSTOR, ArtStor, OCLC, EBSCO, LexisNexis, ProQuest, etc.

There are also a number of service providers in UK Academia who have a remit to provide long-term, stable services, including for instance Edina/EPCC, STFC, etc.

The management of the collections under the umbrella of a management organisation can provide synergies with related collections, and the economies of scale ensures that both the collection management and the delivery is maintained.

Irrespective of whether the management organisation is commercial or non-profit, it typically requires a subscription from participating organisations, but in many cases the academic institutions will have already have such a subscription so the cost is covered by the University Library. Subscription costs vary widely, but JSTOR — https://about.jstor.org — for instance covers many fields of knowledge at a very moderate cost (based on the Carnegie Classifications) — https://carnegieclassifications.iu.edu/ — and provides access to both published and primary source material.

RECOMMENDATIONS

From our review of existing repositories, our discussions with key members of the Materials Science community, and independent research and experience, we see three main options to pursue:

OPTION 1 - DEVELOP AN AGGREGATOR/INDEX/CATALOGUE SERVICE

This is the route taken by many of the most effective and useful services in academia and the general information access world, ranging from Google, through CrossRef's DOI index, JSTOR's historically deep collections, etc. There are many reasons for this being a successful model.

- **From the content supply** side, the owners are responsible for storage and control of their content. Owners see increased access to their content because it is more discoverable and links between records in the index can further improve discoverability. The discovery tools are also typically better than the individual suppliers could provide in a standalone site, and source organisations may not even need to provide a standalone site — essentially just acting as a storage and archive site. They can also use aggregator-provided embargo, access control, quotas, digital signatures, and other content management and security services.

  For The Royce the primary contributors to the index are likely to be individual researchers, who will submit datasets for inclusion in the index. The data *per se* is likely be stored in an institutional repository or other mandated repository, but the index can provide tools for the generation of quality metadata in the categories of material composition, process, provenance, affiliation and funding, license and usage conditions, testing and approval for applications, publications, etc.

  Over time other contributors will likely include academic and similar organisations with policy requirements for storage e.g. in institutional repositories. This is the same as for published works, and is generally accepted as both acceptable and desirable.

  Another category is commercial research and development organisations that participate in the research community, but may have confidentiality requirements, and they will need provisions for escrow, embargo, and restrictions on metadata displayed, e.g. the metadata may be used in search, but may not be displayed. These needs should be clarified during detailed requirements capture.

  In any case, policies regarding the metadata, i.e. catalogue entries, in the areas of quality, guarantees and completeness, should be present and should evolve as experience is gained and the catalogue matures.

- **On the demand side** users have easier access to wider and more diverse content in one place, they have a common set of discovery and access tools and may be able to benefit from subscription bundling if some sources are paywalled. Ideally the index should look like a repository, allowing access to datasets in as seamless a way as possible, with authentication via integration with Eduroam, Athena, Shibboleth, Kerberos, etc. And, of course, users will also be contributors, so roles as consumers and providers must be supported.

- **The Aggregator,** or index owner, focusses on digital curatorship and collection management of the *metadata, and the user experience. C*rucially aggregators are symbiotic with content providers, not competitors. The aggregator deals with such issues as stable identifiers for content, value-added services such as comments sections, user ratings, 'see-also' links, alerts and rules, whether or not the content source provides such services. The consistent experience they provide across the full range of content, irrespective of where it originates, increases researcher productivity, content accessibility and also ultimately leads to the generation of new knowledge and insights that emerge from the aggregation.

If The Royce decides to develop an indexing service, the computational resource requirements for the index *per se* would be fairly lightweight, as the actual datasets are held elsewhere, such as in institutional repositories, local drive stores, etc. It is likely that one would want to host the actual index on a service (e.g. EDINA) which already has the necessary experience, front-end and search capabilities, network and security infrastructure, and user support in place. One could also explore commercial providers such as EBSCO and ProQuest, but that will throw up paywalls and other access issues.

Subcontracting the infrastructure and front-end (UI/UX) would allow The Royce to focus on the Materials Science-specific elements such as locating data sources, developing ontologies and thesauruses, developing policies for metadata, access control, escrow and embargo, meta-data schemas, normalisation, assessment and quality labelling, keyword generation and other discovery aids.

At a technical level the challenges come at the architectural stage, ensuring that the system is expandable across domains, maintainable, capable of accommodating a variety of policies in security, access controls, integration and monitoring of new data sources/repositories, extending ontologies and metadata schemas, etc.

The core abstractions in the architecture — the generalisations of the various functions and data — should encompass the conceivable use cases, and effort spent here will help ensure that the system can readily expand with fewer reworks.

Finally, a key element of any operational system is the curation of the content, and where the data sources are not long-established and stable, it is even more essential that the personnel in the curation role are highly proactive in identifying and recruiting data sources, monitoring them and intervening to ensure that they remain connected or that their data content remains available in the case the source no longer provides it.

OPTION 2 - WORK WITH A THIRD PARTY TO ADD-IN TO THEIR REPOSITORY

Our recommended starting point would be to commission an index (catalogue) of resources as discussed in Option 1 above. This recognises that in many cases the datasets and information resources already have a natural home due to institutional policies, contractual requirements, etc. That being said, in a field such as materials science there are datasets that would benefit from being homed in a facility with specialised discovery and processing tools, or where there are synergies with other datasets held there. Our recommendation with respect to providing repository facilities is that it should be assessed at some later time. i.e. Is there is sufficient need for a storage and processing facility for specific classes of datasets of interest and

if so, identify if there is an existing repository that meets, or could be funded to meet, the requirements for such data. An example is NOMAD for simulation data, which is currently exploring expanding its remit to include observational data and might provide a good partner.

Our recommendation is that that provision of a primary repository should be limited to specialised classes of dataset where a demonstrable need arises.

A secondary recommendation is that consideration should be given at a later stage to provisioning of a storage only depository for datasets where the original source is proposing to remove (de-accession) the dataset or is in danger of failing completely (c.f. portico.org for journals and NOMAD's 10-year retention limit).

That being said, the feedback received during the interview sessions was that the observational data was often so large that it would have to be transferred via physical media, and in any case the description (metadata) would need to be so detailed that the data was not likely to be useful without the lab notebook detailing the exact process taken to prepare the samples. It is therefore not obvious that the datasets are useful for reproducible science in isolation without interaction with the creator. For this reason we suggest that the creation of a bonafide repository needs to be considered cautiously.

N.B. Conversely, it was also reported during the sessions that there was a little availability of generic datasets of materials that could be analysed in any context, where the sample preparation process need not be available. Independently of these recommendations there should be a discussion on the commissioning of generally useful generic datasets.

OPTION 3 - BUILD AND DEPLOY A NEW REPOSITORY

We include this for completeness, but in our opinion this is a sub-optimal and risky solution. The internet is littered with abandoned academic websites/repositories that were funded as part of a project and then pretty much abandoned. Unless there is a compelling need, strong national level support and long-term commitment and funding, then the likelihood is that a new repository will not provide sufficient value to grow, expand its breadth or gain support from the academic community. You may build it, but they may not come.

Whilst one may wish to consider other options for development, the three chosen represent our view of the extreme cases one might consider, with Option 3 as the least desirable option. Our reasoning for this is simple, repositories are generally established as either subject matter specific or institutional. As we have looked across the different operating models and longevity of repositories, those that reside as a core part of the organisation, i.e. are institutional, are those that have longevity. Subject-specific repositories frequently have a life-time closely linked to project funding. As such our strong recommendation is to follow Option 1. We choose Option 1 over Option 2 for fairly simple reasons. Take NOMAD as an example of a standalone repository , which has developed a solid foundation for capturing and making available computational materials data. It has been operating since 2015, indicating some longevity, but it is project-funded, which is a risk to that longevity. On a positive note, the underlying technologies are open source and maintained. They offer a stand-alone instance for local installation (OASIS), that is appealing in terms of being able to begin a new project by building on their efforts. However, there is a concern, and history bears this out, that you would be dependent upon the NOMAD team for enhancements to the core system. In the documentation, NOMAD indicates that not all functionality and tools available in the OASIS version that are available in the NOMAD

version. This ultimately means that, and notwithstanding the longevity issue, The Royce team would have their own development restricted in the medium- to long-term by the rate of progress of the core NOMAD team.

In addition to this, and considering the repository side of the architecture, both the partner organisations to Royce will have policies for storing research data in their own institutional repositories, and there are existing projects linked to The Royce (Manchester and Sheffield) already operating materials data repositories. Given this, and the inherent long-term nature of repositories, it would be more efficient for The Royce to consider acting as an Aggregator (cataloguing) organisation by developing a core indexing services linking to multiple repositories and data stores. This is not so dissimilar to the CHiMaD/NIST Materials Data Facility.

In closing this section, our strong recommendation is to consider Option1 as the best candidate as it provides;

- The greatest flexibility for The Royce,
- it will not enforce the use of 'yet another repository' with the user-base,
- It is a well-trodden path for development,
- It will provide options to integrate a storage service in the future if that is required.

NEXT STEPS

Assuming that Option 1 is the desired next step, the following development activities will need to be considered:

Establish a team:

This will consist of both a Management Panel and Operational Team. We would suggest that the Operational Team is lead by an individual with a track record of developing complex digital research services.

The Management Panel should consist of a focussed group of people with a mixed background in the public and private sectors, in areas related to Materials Science, with a keen appreciation of the value of access to Materials Science Data in accelerating discovery and enabling the UK to meet its Grand Challenges. It is very similar to establishing a board of Trustees, which usually comprises people with knowledge and experience in the area. It could be driven by key researchers with an interest in this project, but should be guided and managed by The Royce as a core project. It will have clear terms-of-reference to govern its activity and support The Royce management and Board in delivering this as a project for the long-term benefit of the partners.

The Operations Team will comprise people with expertise in:

- Front end User experience/User interaction (UX/UI) designers and developers
- Backend management (e.g. DevOps)
- Ontology & Schema development expertise
- Services development team (develop future services based on user feedback, which may include improved search, data publishing tools, etc)
- Curation Services (manages data policies, preservation strategies, monitoring 3rd party asset availability, index organisation and hygiene functions)

- User engagement and training with a marketing function
- User Support

The Operational Team Lead will:

Report to the Management Panel

- Manage the day-to-day activities and priorities of the Operational Team
- Be responsible for working with third party repository owners to put agreements in place to allow The Royce to index content
- Execute the Plan for development, on-going operations and user recruitment for the Service.
- Manage engagement with user groups to solicit feedback on the Service
- Manage contract with service providers (e.g. storage services)
- Develop a long-term strategy for the funding and growth of the service

The Royce will need to establish this type of team to manage a live service. We have included this to ensure that when embarking on this endeavour, there is an understanding of the types of roles that will be required, and some indication of a management structure.

With respect to immediate next steps to scope out in greater detail what would be required, the list below provides detail. This assumes that The Royce would be considering creation of a new service to support the wider Materials Research Community, and would not begin by looking to existing efforts on Manchester and Sheffield as platforms to build upon. Whatever the case, it is clear that those people involved in the existing repository projects, should be considered strongly as part of the inputs to the future project.

The following list of activities will require multiple teams to be in place to fulfil all parts of this set of activities. Each activity should be run and signed off by The Royce to ensure that there is consensus during a first build prototype, which can proceed efficiently and effectively, e.g. the requirements capture needs to be fixed for prototype development. This does not mean that the first prototype is the final service, but it must be designed and built around a minimum viable product (MVP) and then tested.

Technology Requirements Capture:
Detailed documentation will need to be developed to establish the range of services and features required for the first MVP. Across our discussions we have identified the following:

- **Escrow/Embargo** to provide contributors with the control over when and what data they share. This will range from personal contact information to ensure users can be contacted, to which of the uploaded data/ metadata fields are available for search, and which are available for search and publishing to a results set.
- **Access control** to ensure that users activities are logged and secured, that only they are able to modify their own assets. User roles.
- Authentication services integration.
- **Search.** This will be developed iterative during the project. Whilst there will be a focus upon speed in returning results from a UX perspective, developing the search tools over time will also occur, particularly as the ontologies develop and new data source types are added.

- Adding stable handles for reference (e.g. DOIs)
- **Metadata**. Metadata slices and Minimum Requirements for each slice (e.g. generation of quality metadata in the categories of material composition, material processing, provenance, affiliation and funding, license and usage conditions, testing and approval for applications, publications)
- **Data Collection**. Ingestion, metadata collection and (automated) extraction tools. See next bullet point.
- Linking and access to datasets (internal and external). For example, which Institutional Repositories will be accessed, how, what are the data available? Additional repositories such as that used by the LightForm Project (Zenodo), other cloud providers (Google, DropBox, AWS, etc). Agreements will be required with the owners of repositories that are being indexed. Services will need to be developed to both interoperate with data sources, enable extraction of metadata, and an audit tool for the curation team to assess record quality (i.e. sources validation, coverage and accuracy of submitted data (including compliances with data standards and ontologies), and monitor changes in the third party repository where both original and new data are deposited.
- **Interfaces.** Good UX/UI designs will need to be implemented. There are 3 types of interface that will be used by three different types of user.
    - (1) For users posting data, a relatively simplified workflow will be required (examples are already available from services reviewed in this report, e.g. NOMAD),
    - (2) Interfaces for managing the indexes of data from linked Repositories, which will be used predominantly by the curation team, and
    - (3) search and discovery tools interface for all users. Part of the long-term capability will be a continuous review of UX/UI suitability through user feedback (support groups, blogs and usage analytics).
- **Backend solutions.** These will be determined by the specification of the service to be built, but will need to consider the longer-term requirement of curation. Whilst the precise nature of the technology will be defined to some extent by any hosting partner for this service (e.g. Edina are a good example of a long-term service provider), the underlying database technology candidates are PostgreSQL and other relational-database management systems (RDBMS), NoSQL solutions such as Cassandra, ElasticSearch etc.
- **User Workflows** for the interaction with the service, as opposed to the UX, will need to be designed. Some of this may have already been developed in a hosting partner organisation, e.g. security, access control, etc. This may act as a good starting point for design.
- **Curation roles.** Will be at the heart of the operational management of the service in respect of data organisation, quality, availability of the service, updates and maintenance. The roles are diverse and will require a small team with responsibilities across access, security and privacy, policy development and monitoring, metadata schema compliance, data ontologies, quality control, risk management and planning, preservation strategies and policies, and engagement with third party repository owners whose data are being indexed. They will work closely with DevOps, Users (data creators and data users; researchers), subject-matter experts, and front end developers.

An indicative Gantt chart to show the sequencing of these activities is provided (Fig. 1). There is currently no time element associated with this chart and should be taken as a view of activities and their sequencing. That said, we would estimate the following timings:

Task 1 should take 8-10 months to complete (allowing for requirements to interact with users for information and feedback).

Task 2 may initially take time to secure agreement with repository owners to index their content and enable access, but an allocation of 3 months would be appropriate.

Task 3 is dependent upon prior work in earlier Tasks, but should not be a long and protracted selection process based on detailed technical requirements captured. Additionally, the solution must also ensure that it is sufficiently flexible to allow future developments/enhancements. Selecting a host for the service that has prior experience may positively impact on Task 4.

Task 4 is usually a straightforward task once the prior tasks have been embarked upon, but it may may be that a hosting partner will have an influence on this element.

Task 5 is a critical piece of review as it will essentially establish the scale and scope of the project. It will be very much dependent upon The Royce's longer-term aspirations and ability to commit to developing and operating a service for the wider community over time. In any case, this is central to the project and will need to be established in detail very early on in the Service specification phase.

Task 6 is to develop a detailed specification and work-plan for developing the agreed first prototype as a working MVP. Agreeing the specification and associated work plan will determine both the funding requirement, staffing needs and the speed at which the development can occur. At this stage, it is not

Figure 1. Programme Gantt Chart.
Indicative programme showing the path to pilot prototype launch with dependencies. No time frame is indicated.

appropriate to fix a type of development programme (e.g. Agile), as it will be dependent on being able to accurately define the specification and not to deviate from the associated build plan. The specification should enable the service to adhere to FAIR data principles and be built in well-supported, open source technologies to ensure that the service itself does not fall victim to technology obsolescence, not does it struggle to find rare expertise to maintain it.

Task 7 is about putting in place the structures to oversea the development and long-term operation of the service. It needs to be something that is able to engage with Senior Management of The Royce, and be supported at Board Level as a core project. This will be heavily dependent upon the ability to secure funding for hiring the team and development. This should not necessarily sit outside of the The Royce, but should consider procuring services for support as it builds its own capability. Strong project management will be critical.

Task 8 is about building and testing the service, garnering feedback from selected early users, and working towards a full launch. Ahead of launch, the teamed hired should be working towards (i) promoting the service, (2) working to secure access to other repositories to build content, (3) generate training materials to onboard users, and (4) develop project KPIs to report to The Royce Management Panel and Board, and to users. They could be KPIs related to amount and quality of content indexed, numbers of user (searches, contributions), status of relationships with third party providers (hosting, repository owners), number of times a dataset has been accessed, citations, etc.Collecting and reporting KPIs will be key not only to demonstrate the value of the service to The Royce Board and stakeholders, but to the wider user community.

Following Full Launch, continuous monitoring of service use, expanding the user-base and allowing researchers to develop their own tools to access and use the data should be made possible. This service should be planned to be a long-term service acting as a trusted directory of Materials Science Data.

This paper was produced by Impact Data Metrics for the Henry Royce Institute and funded by the Engineering and Physical Sciences Research Council (EPSRC).

This report forms part of a suite of complementary roadmapping and landscaping reports designed to stimulate and drive new advanced materials research in the UK:

Materials 4.0: Digitally-enabled materials discovery and manufacturing

Materials for Fusion Power

Materials for End-to-End Hydrogen

Degradation in Structural Materials for Net-Zero

www.royce.ac.uk

30 March 2021

Henry Royce Institute • Royce Hub Building
The University of Manchester • Oxford Road • Manchester • M13 9PL
ROYCE.AC.UK • INFO@ROYCE.AC.UK