Developing a Community Capability Model Framework for data-intensive research

Liz Lyon
UKOLN, University of Bath
Bath BA2 7AY
United Kingdom
+44 1225 386580
e.j.lyon@ukoln.ac.uk

Alexander Ball
UKOLN, University of Bath
Bath BA2 7AY
United Kingdom
+44 1225 386580
a.ball@ukoln.ac.uk

Monica Duke
UKOLN, University of Bath
Bath BA2 7AY
United Kingdom
+44 1225 386580
m.duke@ukoln.ac.uk

Michael Day
UKOLN, University of Bath
Bath BA2 7AY
United Kingdom
+44 1225 383923
m.day@ukoln.ac.uk

ABSTRACT
Researchers across a range of fields have been inspired by the possibilities of data-intensive research. In many cases, however, researchers find themselves unable to take part due to a lack of facilities, insufficient access to data, cultural disincentives, and a range of other impediments. In order to develop a deeper understanding of this, UKOLN, University of Bath and Microsoft Research have been collaborating on developing a Community Capability Model Framework (CCMF) designed to assist institutions, research funding-bodies and researchers to enhance the capability of their communities to perform data-intensive research. This paper explores the rationale for using capability modelling for informing the development of data-intensive research and outlines the main capability factors underlying the current version of the CCMF.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications - *Scientific databases*

General Terms
Management, Measurement, Performance, Design, Economics, Human Factors

Keywords
Data-intensive research, Fourth Paradigm, capability modeling, research data, managing research data

1. INTRODUCTION
Following the publication of *The Fourth Paradigm* [1], researchers across a range of fields have been inspired by the possibilities of data-intensive research, that is, research involving large amounts of data, often combined from many sources across multiple disciplines, and requiring some degree of computational analysis. In many cases, however, researchers find themselves unable to take part due to a lack of facilities, insufficient access to data, cultural disincentives, and a range of other impediments. In order to develop a deeper understanding of this, UKOLN, University of Bath and Microsoft Research have been collaborating on developing a Community Capability Model Framework (CCMF) designed to assist institutions, research funding-bodies and researchers to enhance the capability of their communities to perform data-intensive research by:

- profiling the current readiness or capability of the community;
- indicating priority areas for change and investment, and;
- developing roadmaps for achieving a target state of readiness.

In this paper, we will introduce the current version of the CCMF, outline some of the concepts underlying it and explain how it came to be in its current form.

2. DEFINITIONS
Data-intensive research belongs to what Gray [2] has termed the Fourth Paradigm of science, that is one primarily based on large-scale 'data exploration'. It is typified by workflows where researchers only apply their academic insight to data after an intense period of data collection and processing, with the processing stages dominant. Most 'big-science' disciplines - e.g., high energy physics, astronomy - are inherently data-intensive, while fields like the life sciences and chemistry have been utterly transformed in recent decades by the sheer quantity of data potentially becoming available for analysis [3]. Even the humanities and social sciences are not exempt from this 'data deluge,' e.g. with the emerging interdisciplinary fields of computational social science [4] and 'culturomics' [5].

One of Gray's key insights was that current data infrastructures were largely insufficient to deal with the vast amounts of data being produced [6, 7]. For example, Kolker, *et al*. [8, p. 142] comment that in the life sciences, "existing data

storage resources and tools for analysis and visualization lack integration and can be difficult to disseminate and maintain because the resources (both people and cyberinfrastructure) are not organized to handle them."

The CCMF is intended to provide a framework for analysing the capacity of communities - through institutions, research funding-bodies and researchers - to deal with data-intensive research. For the purposes of the CCMF, the following characteristics are necessary indicators of data-intensive research:

a) The research typically involves intense computational analysis of data.

b) The research typically involves analysis of large quantities of data, that is, more data than a research team could reasonably be expected to review without software assistance.

Also, if research involves combining data from several different sources, where the different source datasets have been collected according to different principles, methods and models, and for a primary purpose other than the current one, then it is likely to be classed as data-intensive research.

In terms of the CCMF, a *community* is broadly understood to be a set of people who share a particular location within the structure of an institution or society in general. Communities typically engage in both common and collective activities, and develop shared values, vocabularies, strategies and tactics [9]. In the particular case of academia, the term 'community' can apply at several different granularities: from the set of all academics and researchers, to disciplines such as physics or chemistry, or to narrow sub-disciplines such as organic crystallography [10, section 2.4.1]. It can also apply to the academics and researchers within a particular institution or department, or those working on a common project. In the context of the CCMF, the communities we are most interested in modelling are those defined by a discipline, a sub-discipline, or an institution.

3. CAPABILITY MODELS

Capability models are widely used by industry to help identify key business competencies and activities, helping to determine whether, how easily, and how well a given organization or community would be able, in theory and in practice, to accomplish a given task. The project team looked at a range of existing capability models in order to inform the development of CCMF, amongst them the Capability Maturity Model for Software and the Cornell Maturity Model for digital preservation, both of which have been used to explore data management requirements.

3.1 Capability Maturity Model for Software

A particularly influential capability model has been the Capability Maturity Model for Software (CMM) developed by the Software Engineering Institute at Carnegie Mellon University. This is concerned with evaluating the capability of an organisation to develop software on specification, on time and on budget [11]. CMM is a tool that can be used to appraise the current state of an organisation's processes, set targets for how it should be operating, and draw up a roadmap of how to achieve those targets. CMM defines five levels of software process maturity:

1. Initial - software process *ad hoc*, occasionally chaotic
2. Repeatable - basic project management processes established, some process discipline
3. Defined - software process for management and engineering is documented, standardized and integrated
4. Managed - detailed measures of process and quality are collected, software processes understood and controlled
5. Optimizing - incorporating continuous process improvement and innovation

More recently, CMM has been applied to research data management in two independent initiatives. For example, the Australian National Data Service (ANDS) [12] provides descriptions of the five levels of maturity for four key process areas: Institutional policies and procedures; IT Infrastructure; Support Services; Managing Metadata. The ANDS version of the model is much simpler than CMM itself, with narrative descriptions of maturity levels within each process area replacing the sets of key practices and common features. The focus is on higher education institutions, with the four process areas mapping neatly onto groups and services such as senior management, IT support, researcher support or staff development, and the library. The model freely acknowledges that not all organisations will aim to attain Level 5 (optimized) in all areas.

Crowston and Qin [13] take a different approach, focusing on scientific data management within research projects. They interpret the five levels as follows.

1. Data are managed within the project on an *ad hoc* basis, following the intuitions of the project staff.
2. Plans, policies and procedures are in place for data management, but they are peculiar to the project and reactive in nature.
3. The project tailors for itself plans, policies and procedures set up for data management at the discipline, community or institutional level; these plans tend to be pro-active in nature.
4. The project measures the success and effectiveness of its data management to ensure standards are maintained.
5. The project identifies weaknesses in its data management and addresses the defects pro-actively.

In developing their version of the model, Crowston and Qin consulted data management literature to identify key practices in data management, which they grouped into the following four key process areas:

1. Data acquisition, processing and quality assurance (3 practices)
2. Data description and representation (7 practices, including 'Develop and apply metadata specifications and schemas', 'Design mechanisms to link datasets with publications', 'Ensure interoperability with data and metadata standards')
3. Data dissemination (4 practices, including 'Encourage sharing', 'Distribute data')
4. Repository services/preservation (7 practices, including 'Store, backup and secure data', 'Perform data migration', 'Validate data archives')

In addition, they identified several generic practices that closely resembled those in the earlier models, for example:

developing policies for data release, sharing, data rights and restrictions, and data curation; identifying staffing needs; developing business models; developing data management tools; training researchers and support staff ; capturing provenance data; developing collaborations and partnerships; assessing impact and enforcing policy.

The use cases for all of these capability models strongly resemble those intended for the CCMF. They provide a clear framework for characterising an organisation or project, and identifying improvements that could be made as well as the order in which they should be tackled. They also provide a reference vocabulary for describing relevant activities and functions, without being overly specific about how these should be carried out or implemented. While CMM is primarily focused on the commercial sector, the version of the model developed by ANDS shows, however, how it can be applied to higher education institutions. Crowston and Qin's model focuses on research projects while also referencing (and having clear implications for) the wider institutional and disciplinary context. Indeed, perhaps the most important difference to reconcile between these models and what is required for the CCMF is that they again admit only one target state to which organisations should aspire, with the possible exception of the ANDS model; in contrast, it would be difficult to find a single generic description that could apply to all successful forms of data-intensive research.

3.2 Cornell Maturity Model

A slightly different approach to capability modelling was developed in the Cornell Maturity Model used to analyse the type of response given by higher education institutions to the challenges of digital preservation. Kenney and McGovern [14, 15] present a distinctive five-stage maturity model:

- Acknowledge. The institution recognises it must perform some degree of digital preservation.
- Act. The institution instigates digital preservation projects.
- Consolidate. The institution embeds digital preservation as ongoing programmes.
- Institutionalise. The institution unifies the various digital preservation activities into a single programme.
- Externalise. The institution collaborates with others to achieve economies of scale and increased digital preservation capability.

In the early expressions of the Cornell model, key indicators for each stage were described along the three dimensions of policy and planning, technological infrastructure, and content and use. These dimensions were later changed to organisational infrastructure, technological infrastructure, and resources, with a corresponding new set of key indicators. To emphasise that organisations should develop in each of the dimensions in parallel, but that the digital preservation capability can still be stable with uneven development, they became known as the three legs of a digital preservation Three-Legged Stool, with legs for organization, technology and resources.

The Cornell model was further developed by the JISC-funded AIDA Project into a scorecard-based tool for benchmarking the current state of digital asset management within institutions or departments. AIDA expanded and formalised the indicators within each leg, arriving at eleven metrics in each of the organisation and technology legs, and nine metrics within the resources leg. While AIDA was intended as a self-assessment toolkit, the AIDA Project Team provided a service for assessing completed scorecards to determine an overall picture of institutional readiness, recommend actions for increasing readiness, and provide guidance on digital asset management issues.

The AIDA scorecard provided by the Project Team was in the form of a Microsoft Word document with form controls, with analysis performed on an accompanying Excel spreadsheet. The process of performing the benchmarking exercise itself, though, was left up to the individual to plan. Sensing a need, the UK Digital Curation Centre (DCC) applied its experience from developing the tools that supported DRAMBORA and the Digital Asset Framework (DAF) to produce a Web-based tool allowing a team of contributors to collaborate on an AIDA-style self-assessment. This tool, known as CARDIO [16], uses a very similar set of metrics ('statements') to those developed by AIDA, but has a specific emphasis on research data and can be used at multiple levels of organizational granularity (project, department, institution).

The use cases for this model – assessing the current state of readiness of an institution and identifying priorities for development – again resonate strongly with those for the CCMF. Just as the CCMF should be applicable to researchers, institutions and funding bodies, the Three-Legged Stool can be applied at several different granularities. The notion of having broad, abstract dimensions measured according to specific, concrete metrics is a useful one. Once more, though, the model considers only one correct route from nil readiness to complete readiness through each leg, and through each metric within each leg. The CCMF, by contrast, needs to model several types of community capability and - by implication - several different 'routes' to achieving capability.

4. CCMF CAPABILITY FACTORS

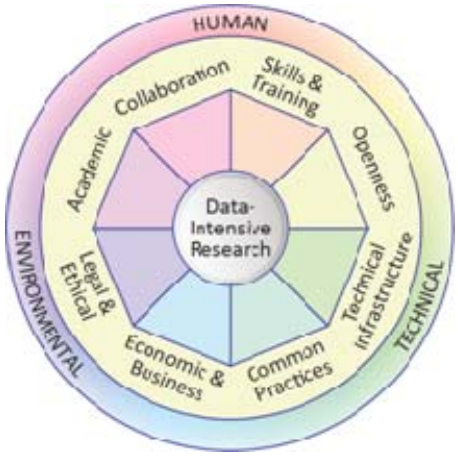

Figure 1: Community Capability Model Framework

We propose a Community Capability Model Framework for data-intensive research comprising eight capability factors representing human, technical and environmental issues (Figure 1). Within each factor are a series of community characteristics that we feel are relevant for determining the capability or readiness of that community to perform data-intensive research. In this section, we

will outline the eight capability factors that make-up the CCMF and comment on some of the characteristics associated with each one. The CCMF consultation draft [17] provides additional detail on all of these, including:

- an identification of the community characteristics associated with each factor, including indications of how each characteristic could be 'measured' for the purposes of analysis and comparison;
- one or more exemplars demonstrating how the alternatives should be interpreted, and;
- brief commentary explaining the relevance of the characteristic for determining capability, and how the project team's thinking has been shaped by the literature and by discussions with the community. These discussions took place in a series of five workshops held between September 2011 and February 2012 in the UK, US, Sweden and Australia.

4.1 Collaboration

The working relationships that are formed during research have a strong bearing on the types of research that can be performed. Collaborations can be informal or semi-formal, or can be rigorously controlled, managed and enforced through contracts and agreements. Collaboration can be organised within a discipline, between two or more disciplines, with organizations outside the research sector, and with the general public.

4.1.1 Collaboration within the discipline/sector

The level of collaboration within a discipline can range from almost none (sometimes characterised as the lone researcher) to extremely large, well-organised national or international consortia. In practice, however, perhaps most disciplinary collaboration is focused on a particular research group or groups. For example, bioinformatics and neuroinformatics are dominated by small teams, with relatively few large-scale contributors. By contrast, big science disciplines like high energy physics and astronomy are typically organised in projects at international scale.

It is recognised that individual researchers can move along the spectrum as their career progresses, e.g. first working alone on an idea or hypothesis, exposing it gradually to colleagues and gaining collaborators from the research group and, at a later stage, the wider community.

4.1.2 Collaboration/ interaction across disciplines

Interdisciplinary collaborations follow the same broad pattern as those within disciplines. Some disciplines will have next to no interaction with others while others will have forged formal collaborations over relatively long periods of time.

Interdisciplinarity is one response to the perceived over-specialisation of research disciplines, and can be encouraged in institutional or national contexts through the creation of matrix structures like joint research centres or faculty appointments [18, pp. 173-4]. Data-intensive research will tend towards the interdisciplinary, not least because it requires the input of computational specialists. There are many potential impediments to interdisciplinary collaboration, not least epistemic barriers based upon what Jacobs and Frickel [19, p. 47] describe as "incompatible styles of thought, research traditions, techniques, and language that are difficult to translate across disciplinary domains."

4.1.3 Collaboration/ interaction across sectors

Researchers will sometimes need to collaborate across sector boundaries, e.g. with industry, equipment suppliers, media, professional bodies or public sector organisations. The types of organization suitable for collaboration will vary quite widely, anf might include: pharmaceutical and biotechnology companies (in medicine and the life sciences), natural history museums (in biodiversity, ecology and palaeontology), or the digital content industries (e.g., Google Book Search for culturonomics).

4.1.4 Collaboration with the public

There is a growing interest in public engagement with research. This is particularly strong in the life sciences, where some funding bodies (e.g., medical research charities) are keen to involve patients in things like reviewing grant proposals. In fields as divergent as astronomy (GalaxyZoo) and papyrology (Ancient Lives), members of the public are being encouraged to contribute directly to some aspects of the research process.

4.2 Skills and training

The capability of a community to perform data-intensive research is strongly influenced by the individual capabilities of its members, and the capacity that results from the combination and multiplication of these capabilities. Community capability can therefore be enhanced by training members in the relevant skills. This training is most effective when it is fully embedded as part of the early education and continuing professional development of researchers.

4.2.1 Skill sets

The capability of a community to perform data-intensive research is strongly influenced by the individual capabilities of its members, and the capacity that results from the combination and multiplication of these capabilities. Community capability can therefore be enhanced by training members in the relevant skills. This training is most effective when it is fully embedded as part of the early education and continuing professional development of researchers.

4.2.2 Pervasiveness of training

There is much variation across disciplines, institutions and degree programmes in the provision of training. Some UK research funding bodies have established Doctoral Training Centres to develop and deliver training programmes for their disciplinary communities. JISC has funded training materials that target particular disciplines e.g. psychology. At some institutions - including the University of Bath - support services like subject liaison librarians and IT services are beginning to develop a range of training programmes for researchers, covering topics such as data management planning. The UK Digital Curation Centre has delivered training modules on a regional basis as part of its Regional Roadshow Programme, while national data centres such as the ESDS (in the UK) and ICPSR (in the US) run workshops on data management.

4.3 Openness

Historically, scientific progress has been driven forward by the open communication of research methods and results. More generally, the principle of openness can be applied at different levels: from openness in communicating the plans for research and ongoing progress whilst the research is undertaken, to opening up the published literature to a wider audience. Driven by concerns for improving the validation, reproducibility and

reusability of research, the last decade has also seen calls for opening up the data and other details of methodologies employed, alongside final results and conclusions, for scrutiny and re-use by the wider community, a process that is considered by some to add value to research.

4.3.1 Openness in the course of research

This characteristic describes whether researchers choose to communicate information about their research whilst it is still ongoing, the extent to which they make their plans and intermediate results known, and the mechanisms they use to achieve openness. Such openness makes an informal variety of early peer review possible, which in the long term may result in more interoperable data and therefore more opportunities for data-intensive research.

4.3.2 Openness of published literature

The body of published literature can be available under different conditions – some literature is available only through payment agreements; sometimes only the description of the literature (metadata) is accessible, whilst at the other extreme some communities have embraced the practice of sharing of all the published literature through archives freely available to all. The openness or otherwise of a publication may depend on its type (journal paper, conference paper, thesis), or readers may need to make specific personal requests in order to gain access. Providing open access to published literature may make it easier for potential re-users to locate suitable data.

4.3.3 Openness of data

There is wide variation in the openness of data. In some disciplines, e.g. astronomy, proteomics and philology, data is routinely published openly, sometimes after a period of exclusive use. In others, there is no tradition of data sharing. For example, O'Donoghue, *et al*. [20] note the unevenness of availability of biological data, with the two extremes exemplified by PDB, which contains almost all experimentally determined structures, and image data from high throughput experiments, where there is little data integration and 'most of these data are never made publicly available'.

Treloar [21] presents a model of data openness with the following three categories:

1. Private research domain. Typically access is tightly controlled and restricted to a core team within a single institution. Technological platforms such as laboratory information management systems or research management systems are used.
2. Shared research domain. This is where some, but not all, the data is shared by the core team with other colleagues, often outside the home institution.
3. Public domain. Data is published so that (with a few exceptions) anyone can gain access to it. Institutional repositories may be used to provide this access. Typically the data will be given a persistent identifier, and the associated metadata will be fixed.

4.3.4 Openness of methodologies/workflows

Releasing data alone may not be sufficient to replicate results and findings. Details of methodologies and workflows which allow other researchers to reproduce the workings and methods of other groups may be required. This characteristic describes the practice of sharing information regarding the processes employed, either as descriptions or in executable forms, so that one researcher can apply the same methods either to the same dataset or perhaps to alternative data or applications.

4.3.5 Reuse of existing data

This characteristic focuses on the attitudes and practices of using data sets generated by other researchers. Researchers may be open to regularly using data shared by others, but they may only trust specific sources. Data sets obtained from the community can be processed in different ways – data can be aggregated, re-analysed under the original conditions or mined to generate new insights.

4.4 Technical infrastructure

The technical infrastructure that supports research comprises tools and services that are used at different the stages of the research life cycle. This capability factor describes categories of tools and services that meet user needs across various activities.

4.4.1 Computational tools and algorithms

Computational tools and algorithms form the backbone of most data-intensive research workflows. If such tools under perform, it places a hard limit on what research can be conducted.

4.4.2 Tool support for data capture and processing

Tools that support data capture and processing often make assumptions about the formats in which the data is stored and processed. The extent to which the tools support formats that are more widely supported by other tools may determine whether data can be shared, understood, processed and re-used within the wider technical environment. When the tools support open or agreed formats or the interchange of data in different formats, tool interoperability increases.

4.4.3 Data storage

Data storage needs to grow as data volumes increase, but requirements may also be defined by the data type. Such requirements may involve issues of physical location, performance, access control and security, scalability, reliability, and speed as well as capacity. For example, in some communities the storage of clinical data must adhere to the ISO/IEC 27000 series of information security standards. Data storage can be organised locally, nationally or globally. Interactions with data storage are required by several of the other tool categories, such as data capture and processing tools, discovery services and curation and preservation services.

4.4.4 Support for curation and preservation

The relative importance of the tools that enhance contemporary usefulness of data and those that aid its long-term preservation varies between disciplines. For disciplines reliant on non-replicable observations, good preservation tools help to maintain stocks of data for future data-intensive research.

4.4.5 Data discovery and access

Data discovery and access is currently problematic because different types of catalogues do not integrate well and there is no standard way to publish them, and no easy way to federate them for cross-discovery. Other challenges exist at the semantic level [22, 23]. One measure suggested would to see how far a community might be from agreeing standards.

4.4.6 Integration and collaboration platforms

Integration and collaboration tools may help researchers manage their workflows and interactions more efficiently, increasing their capacity for data-intensive research.

4.4.7 Visualisations and representations

Visualisation tools are extremely important for data-intensive science. However, the current range of visualisation tools tends to be fragmented and not necessarily optimized for the scales of data becoming available [20].

4.4.8 Platforms for citizen science

Citizen science platforms provide infrastructure that enables non-specialists to participate and collaborate in the research process. Whilst the platforms can be developed within a specific project they can then be redeployed to meet the need of other communities.

4.5 Common practices

This capability factor describes community practices that have produced standards, whether by design or *de facto*. The quantity of standards in a particular discipline is not necessarily a measure of its capability. In some cases, standards may actually hold back progress, especially where they are poorly supported by software or where competing standards effectively act as data silos. It is the quality of data standards that is important, specifically whether they promote and enable the re-use and combination of data. While convergence on a *de facto* standard can happen organically, designed standards typically need to be driven either by influential organisations at a national or international level, or else by a dedicated and enthusiastic association of individuals within a community.

4.5.1 Data formats

These are formats that describe how data is encoded and stored, and facilitate data exchange.

4.5.2 Data collection methods

Data collection methods can also be standardised and shared. Methods are varied depending on the activity within which collection is undertaken. Data collection activities include observational collection, instrumental collection requiring calibration, survey data, sensor data and performance data.

4.5.3 Processing workflows

If data has been processed according to standard and accepted workflows, it is more likely to considered for reuse by other researchers.

4.5.4 Data packaging and transfer protocols

Agreed standards for data packaging and transfer ease the transport of data between creators, archives and the re-users of data.

4.5.5 Data description

Data description standards are used to make data re-usable by providing metadata that describes different aspects of the data. Whilst some disciplines have adopted description schemes that become widely used, other schemes are at earlier stages of adoption and have not yet fulfilled the promise of data interoperation and reusability that they are intended to facilitate. Schemes can be aimed at a generic level or be specialised with discipline-specific fields.

4.5.6 Vocabularies, semantics, ontologies

Vocabularies, semantics and ontologies are also used by communities to exchange information and data, and attempt to capture the knowledge, concepts and terminologies within the discipline in a standardised agreed format. Some are adopted within specialised communities, whilst others find their place as a bridge between communities. Different models for how these standards are agreed and maintained can be described, and their progression or maturity follows a trajectory from proposal and specification to standardisation by recognised bodies.

4.5.7 Data identifiers

Data identifiers are developed to provide unique and unambiguous methods to refer to or access research objects. They may serve the purposes of identification and location. The objects may be literature, chemical or biological entities, or entries in databases.

4.5.8 Stable, documented APIs

Where data repositories and data processing services provide APIs, it opens up the possibilities for automated workflows and thereby increases the scale at which research can be performed.

4.6 Economic and business models

Moving into data-intensive research requires some degree of investment, and it is therefore important to consider how this might be funded and the business case for making the move. Disciplinary differences are important here: the business case will be easier to make where it is important to publish quickly and generate many research papers from a single investment, and harder where the emphasis is on careful and considered weighing of evidence.

4.6.1 Funding models for research and infrastructure

There are many thematic perspectives to consider here including scholarly communication and data publishing models, approaches to data curation and preservation, network-level infrastructure, through to capacity-building programmes. The established political and funding landscape in a particular geographical area is strongly influential in determining the business models in place. In order to realise the full potential global scale of data-intensive research, politico-legal issues and barriers linked to trans-national borders, will need to be overcome.

4.6.2 Public–private partnerships

In communities where it is common for research to be partially or wholly funded by the private sector, the diversity of funding streams may make the research more sustainable, and the research may have greater impact outside academia. At the same time, the research may be contingent on business models and return on investment, and it is less likely that data will be made available for reuse.

4.7 Legal and ethical issues

Quite apart from any cultural barriers that may obstruct data sharing, and thereby restrict the scope for data-intensive research, in some cases there may be ethical reasons why certain datasets may not be shared, and legal barriers both to sharing data in the first place and to recombining it for the purposes of data-intensive research. Even in cases where the barriers do not in fact exist, ambiguities and misperceptions of the legal or ethical position may deter risk-averse institutions and researchers from pursuing

such lines of enquiry. It will, therefore, be easier for data-intensive research to flourish where the legal issues surrounding data sharing and reuse are well understood and well managed, and where there are established frameworks for ensuring such research is conducted in an ethical manner.

The following characteristics should be assessed with caution, as the official policies do not always reflect what is actually done by researchers and institutions.

4.7.1 Legal and regulatory frameworks

At issue here are laws that impact on the sharing and reuse of data (most notably intellectual property laws and contract law), as well as relevant policies and regulations adopted by governments, funding bodies, professional societies and other bodies. The benefit of legal and regulatory frameworks for community capability lies in the clarity they provide with respect to the law, so that it is readily apparent whether and how data may be shared and reused. In the UK, such frameworks might, for example, instruct researchers to record the owner of data, to avoid future uncertainty over the contractual arrangements under which the researcher was working. There are several points of failure, though, that must be avoided. No framework will be able to work around firm legal prohibitions. In some US jurisdictions there are limitations on state-based contracts, signing contracts outside of the state, and selling outside the state by state-based institutions. Where the law itself is ambiguous or untested, any framework for managing compliance will necessarily be cautious. More helpful frameworks may build on the firmer parts of the law to allow routes for data sharing and reuse, while more obstructive frameworks might block the possibility entirely. Even where helpful frameworks do exist, researchers must be familiar with them and trust them. Funding bodies, professional societies, governing bodies and regulators play a large part in ensuring adherence to procedures and community norms, but their attitudes may not always be favourable to the needs of data-intensive research.

4.7.2 Management of ethical responsibilities and norms

As with the previous characteristic, the issue here is with clarity. Researchers will feel more confident about releasing sensitive data if there are established and trusted procedures in place for anonymising it, limiting access to it, and so on. There are also ethical issues relating to research quality.

4.8 Academic culture

The community norms that exist for the process of doing research are a key factor in determining the level of support a researcher might expect when moving into data-intensive research. Such a move may be easier where entrepreneurship and innovation are welcomed, and harder where such things are frowned upon. Even more importantly, data-intensive research is most likely to flourish in communities where data is valued highly, where researchers are rewarded for their data contributions, and where high standards are expected of data entering the research record.

4.8.1 Productivity and return on investment

The impact that this characteristic has on community capability is relatively weak but it is still important to recognise. While the metric is couched in terms of timescales and publishing patterns, the underlying feature we are interested in is the character of the research. The rapid-cycle end of the dimension is the natural home for disciplines where the interest is in finding new things: new particles, new molecules, new sequences. The slow-cycle end of the dimension is the natural home for disciplines where the interest is in profound insight, and improved understanding of complex issues. Data-intensive research methods can assist in all these areas of enquiry, but the immediacy of their impact varies. At the rapid-cycle end, it is relatively straightforward to decide which patterns to look for in data, and relatively obvious when an interesting result has been found; in such cases an investment in the means of data-intensive research has a quick pay-off. At the slow-cycle end, it is typically harder to assemble a comprehensive dataset to analyse, and the analytical steps to automate may themselves require debate and justification; in such cases, greater preparation is needed before data-intensive research methods are applied, and once they are it may take some time to reap the benefits.

4.8.2 Entrepreneurship, innovation and risk

The move to a new paradigm of research requires a certain degree of investment, in both time and effort, and there is always a risk that it may not produce interesting results, or that peer reviewers may not accept the new methodology. There is therefore risk to both PIs and funding bodies when it comes to funding such research. In disciplines where risk-taking and innovation are seen in a positive light, this is less of a barrier.

4.8.3 Reward models for researchers

Contributions to data intensive research are made in different ways. Not all these contributions are formally recognised when considering rewards for researchers. Rewards can come in many forms including career advancement, recognition by peers and funding, both for research and for training students and junior researchers. Methods for measuring contributions are also varied, with some measures for example publications being well established. Even within publications, however, there are different ways of recording contribution. Multi-author efforts can credit each contributor. Other categories of contribution encompass software products and sharing of analysed data, such as DNA sequences. Some contributions such as efforts to curate data and make it reusable are notorious for being poorly recognised and rewarded.

4.8.4 Quality and validation frameworks

Even if data is shared, it may not be in a state amenable to reuse, let alone full validation. Unless data is sufficient quality, and provably so, it is of limited use in data-intensive research conducted by other researchers. A community's capability for such research, therefore, is increased where data is available that has been through thorough independent quality checks, and where this data is maintained and integrated with similar data by specialist curators.

5. CONCLUSIONS

The Community Capability Model Framework is a tool for evaluating a community's current readiness to perform data-intensive research, and for identifying areas where changes need to be made to increase capability. This paper has outlined the eight capability factors identified, which deal with human, technical and environmental issues. The detailed CCMF [17] attempts to identify characteristics that can be used to judge community capability.

While the CCMF has been developed with the involvement of a wide range of stakeholders and interested parties, the immediate next step will be to validate it by applying the

framework to a number of research communities. In the longer term we hope to develop tailored versions of the framework for different stakeholders, and to improve the usefulness of the tool as an aid to decision making and planning.

6. ACKNOWLEDGMENTS

The CCMF was developed as part of the Community Capability Model for Data-Intensive Research project, a partnership of Microsoft Research Connections and UKOLN, University of Bath: http://communitymodel.sharepoint.com

The authors would like to thank all those that attended CCMDIR workshops for their participation in the development process and for their comments on earlier drafts of the model.

7. REFERENCES

[1] Hey, T., Tansley, S. and Tolle, K. Eds. 2009. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, Redmond, WA.

[2] Gray, J. 2009. Jim Gray on eScience: a transformed scientific method. In *The fourth paradigm: data-intensive scientific discovery*, T. Hey, S. Tansley and K. Tolle, Eds. Microsoft Research, Redmond, WA, xix–xxxiii.

[3] Hey, T., and Trefethen, A. 2003. The data deluge: an e-science perspective. In *Grid computing: making the global infrastructure a reality*, F. Berman, G. C. Fox, and T. Hey, Eds. Wiley, New York.

[4] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., *et al*.. 2009. Computational Social Science. *Science* 323 (6 Feb), 721-723. DOI= http://dx.doi.org/10.1126/science.1167742

[5] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., *et al*. 2011. Quantitative analysis of culture using millions of digitised books. *Science* 331 (14 Jan), 176-182. DOI= http://dx.doi.org/10.1126/science.1199644

[6] Gray, J., Liu, D. T., Nieto-Santiseban, M., Szalay, A., DeWitt, D. J., and Heber, G. 2005. Scientific data management in the coming decade. *ACM SIGMOD Record* 34, 34-41. DOI= http://dx.doi.org/10.1145/1107499.1107503

[7] Szalay, A., and Blakeley, J. A. 2009. Gray's Laws: database-centric computing in science. In *The fourth paradigm: data-intensive scientific discovery*, T. Hey, S. Tansley and K. Tolle, Eds. Microsoft Research, Redmond, WA, 5-11.

[8] Kolker, E., Stewart, E., and Ozdemir, V. 2012. Opportunities and challenges for the life sciences community. *OMICS: A Journal of Integrative Biology* 16, 138-147. DOI= http://dx.doi.org/10.1089/omi.2011.0152

[9] Agre, P. 1998. Designing genres for new media: social, economic, and political contexts, In *Cybersociety 2.0: revisiting computer-mediated community and technolo*gy, S. G. Jones, Ed. SAGE, Thousand Oaks, CA, 69–99.

[10] Treloar, A. 1998. Hypermedia online publishing: the transformation of the scholarly journal. PhD thesis, Monash University, Melbourne. http://andrew.treloar.net/research/theses/phd/index.shtml

[11] Paulk, M. C., Curtis, B., Chrissis, M. B., and Weber, C. 1993. *Capability maturity model*, Version 1.1. Technical Report, CMU/SEI-93-TR-024 ESC-TR-93-177. Carnegie Mellon University, Software Engineering Institute, Pittsburgh PA. http://www.sei.cmu.edu/reports/93tr024.pdf

[12] Australian National Data Service. 2011. *Research Data Management Framework: Capability Maturity Guide*. ANDS Guides. http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.html

[13] Crowston, K. and Qin, J. 2012. A capability maturity model for scientific data management: evidence from the literature. *Proceedings of the American Society for Information Science and Technology* 48, 1-9. DOI= http://dx.dor.org/10.1002/meet.2011.14504801036

[14] Kenney, A. R., and McGovern, N. Y. 2003. The five organisational stages of digital preservation. In *Digital libraries: a vision for the 21st century*, P. Hodges, M. Sandler, M. Bonn, and J. P. Wilkin, Eds. University of Michigan Scholarly Publishing Office, Ann Arbor, MI. http://hdl.handle.net/2027/spo.bbv9812.0001.001

[15] Kenney, A. R., and McGovern, N. Y. 2005. The three-legged stool: institutional response to digital preservation. 2nd Convocatoria del Coloquio de marzo, Cuba, March. http://www.library.cornell.edu/iris/dpo/docs/Cuba-ark-nym_final.ppt

[16] Digital Curation Centre, CARDIO: http://cardio.dcc.ac.uk/

[17] Lyon, L., Ball, A., Duke, M., and Day, M. 2012. *Community Capability Model Framework* (consultation draft). UKOLN, University of Bath, Bath. http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-v1-0.pdf

[18] National Academy of Sciences. 2004. *Facilitating interdisciplinary research*. National Academies Press, Washington, DC.

[19] Jacobs, J. A., and Frickel, S. 2009. Interdisciplinarity: A Critical Assessment. *Annual Review of Sociology* 35 (2009), 43-65. DOI= http://dx.doi.org/10.1146/annurev-soc-070308-115954

[20] O'Donoghue, S.. I., Gavin, A. -C., Gehlenborg, N., Goodsell, D. S., Hériché, J. K., North, C., *et al*. 2010. Visualizing biological data – now and in the future. *Nature Methods*, 7, S2–S4. DOI= http://dx.doi.org/10.1038/nmeth.f.301

[21] Treloar, A. 2011. Private research, shared research, publication, and the boundary transitions. http://andrew.treloar.net/research/diagrams/data_curation_continuum.pdf

[22] Bowker G. C. 2001, Biodiversity dataversity. *Social Studies of Science* 30, 643-84. DOI= http://dx.doi.org/10.1177/030631200030005001

[23] Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., and Borgman, C. L. 2011. Science friction: data, metadata, and collaboration. *Social Studies of Science* 41, 667-690. DOI= http://dx.doi.org/10.1177/0306312711413314