Introducing the Community Capability Model Framework and White Paper

Alex Ball

14 January 2013

Contents

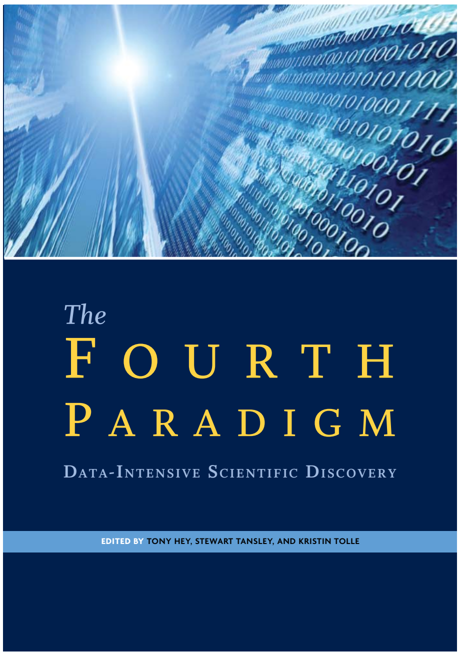
1 Background

1.1 The Fourth Paradigm

The inspiration behind our project comes from the book *The Fourth Paradigm*, published by Microsoft (Fig. 1). This book contains several examples of data-intensive research being carried out in the areas of environmental science, astronomy, biology and medicine, alongside discussions of the technology that makes it possible and the implications for scholarly communication.

Underlying all this is an idea put forward by the late computer scientist Jim Gray that data-intensive research is a fourth paradigm or way of working for science. It's easy to lose focus on what 'data-intensive' means in this context, but I hope this illustration will make it clear (Fig. 2).

In the beginning was the world. If people are interested enough in an aspect of it, they will try to understand it and thereby produce a model *(reveal)*. If the model is straightforward enough they will be able to solve the equations analytically and derive predictions *(reveal)*. They then compare their predictions to what actually happens and either validate or improve their model *(reveal)*. As time goes on, the problems that are left get harder and standards rise, so the models get more complicated, until the equations can't be solved analytically any more, and we need to run simulations or brute force methods using computers *(reveal)*. Once again, the calculations are compared with reality in order to validate or improve the model *(reveal)*.


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Microsoft® Research

“This book presents the first broad look at the rapidly emerging field of data-intensive science, with the goal of influencing the worldwide scientific and computing research communities and inspiring the next generation of scientists.”

http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Figure 1: *The Fourth Paradigm* book

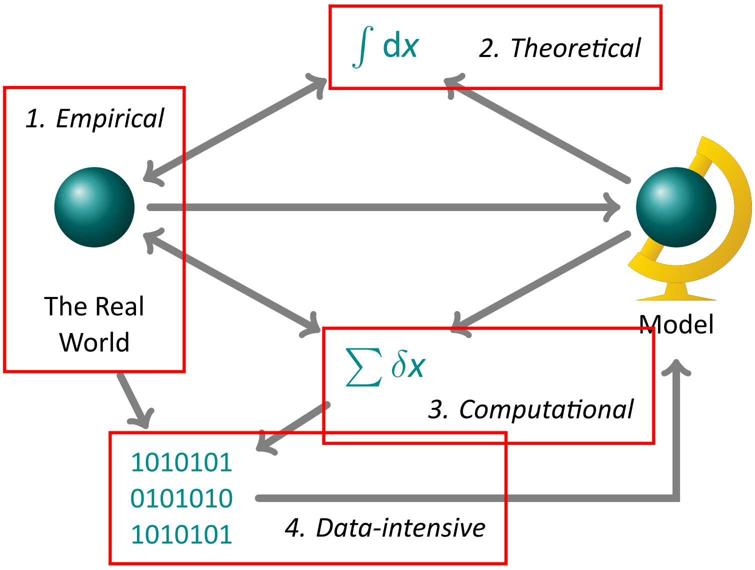


Figure 2: Four paradigms

All the while, these simulations and measurements of the real world are generating vast swathes of data *(reveal)*, and hidden in there are patterns relating to phenomena we don't have models for yet. So there is a new type of activity researchers can perform, which is mining the data for conclusions without the aid of model, though new models may well arise as a result *(reveal)*.

So there *(reveal)* are our four paradigms: *(reveal)* empirical, *(reveal)* theoretical, *(reveal)* computational, and *(reveal)* data-intensive.

The publication of *The Fourth Paradigm* generated a lot of interest. Researchers began approaching Microsoft asking for advice on and support with moving into data-intensive research. But the work that went into the book had not produced a magic formula for cultivating successful data-intensive research within a community. So Microsoft funded the CCMDIR Project to work out, if not a magic formula, then at least a model that could be used to indicate how well data-intensive research might thrive within a given community, and to decide the best ways of improving the situation.

1.2 Project objectives

More specifically, our objectives *(slide)* were to look at data-driven research going on at the moment in different communities, pick out the factors that help it to thrive and determine ways of measuring or categorising those factors. We then wanted to package all that into a compelling Community Capability Model Framework, which could be validated through a series of case studies.

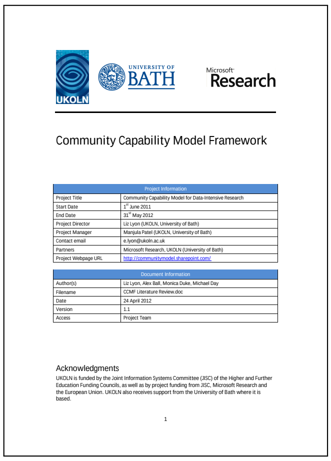
2 The White Paper

After a few iterations, our proposed *Community Capability Model Framework* was published on 24 April last year as a white paper. As you can see, we have high hopes for what it will enable people to do.

The *Community Capability Model* Framework (CCMF) will provide support for:

- Intelligence-gathering
- Decision-making
- Planning
- Investment
- Building capacity
- Building capability
- Knowledge transfer

Liz Lyon



Community Capability Model Framework

Project Information	
Project Title	Community Capability Model for Data-Intensive Research
Start Date	1st June 2011
End Date	31st May 2012
Project Director	Liz Lyon (UKOLN, University of Bath)
Project Manager	Manjula Patel (UKOLN, University of Bath)
Contact email	e.lyon@ukoln.ac.uk
Partners	Microsoft Research, UKOLN (University of Bath)
Project Webpage URL	http://communitymodel.sharepoint.com/

Document Information	
Author(s)	Liz Lyon, Alex Ball, Monica Duke, Michael Day
Filename	CCMF Literature Review.doc
Date	24 April 2012
Version	1.1
Access	Project Team

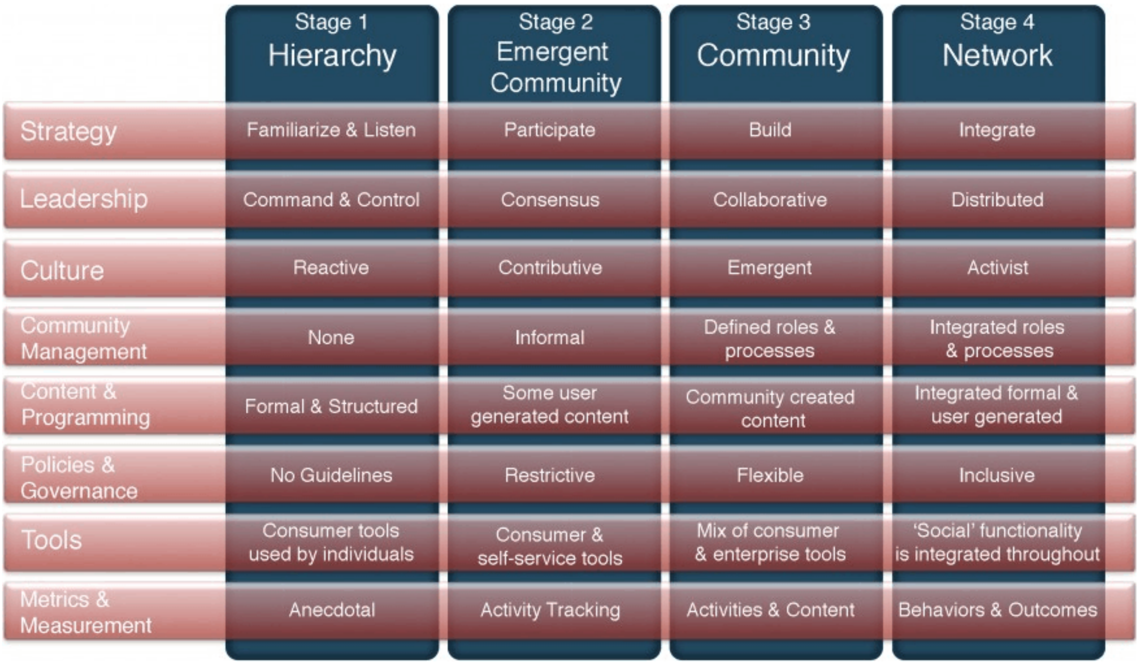
Acknowledgments

UKOLN is funded by the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from JISC, Microsoft Research and the European Union. UKOLN also receives support from the University of Bath where it is based.

1

I should explain that by 'people' I mean 'communities' *(reveal)*, and by 'communities' we mean disciplines, sub-disciplines and super-disciplines, as represented by P.I.s and funders, and also institutions.

By *(reveal)* 'capability model', we mean a model for determining whether, how easily, and how well an agent could, in theory and in practice, accomplish a given task.



	Stage 1 Hierarchy	Stage 2 Emergent Community	Stage 3 Community	Stage 4 Network
Strategy	Familiarize & Listen	Participate	Build	Integrate
Leadership	Command & Control	Consensus	Collaborative	Distributed
Culture	Reactive	Contributive	Emergent	Activist
Community Management	None	Informal	Defined roles & processes	Integrated roles & processes
Content & Programming	Formal & Structured	Some user generated content	Community created content	Integrated formal & user generated
Policies & Governance	No Guidelines	Restrictive	Flexible	Inclusive
Tools	Consumer tools used by individuals	Consumer & self-service tools	Mix of consumer & enterprise tools	'Social' functionality is integrated throughout
Metrics & Measurement	Anecdotal	Activity Tracking	Activities & Content	Behaviors & Outcomes

http://community-roundtable.com/2009/06/the-community-maturity-model/

Figure 3: The Community Maturity Model

2.1 Capability models

Here are some examples of capability models we looked at. The **Community Maturity Model** (Fig. 3) characterizes organizations as belonging to one of four stages according to how they interact with their community, where that means target audience or pool of potential customers and clients.

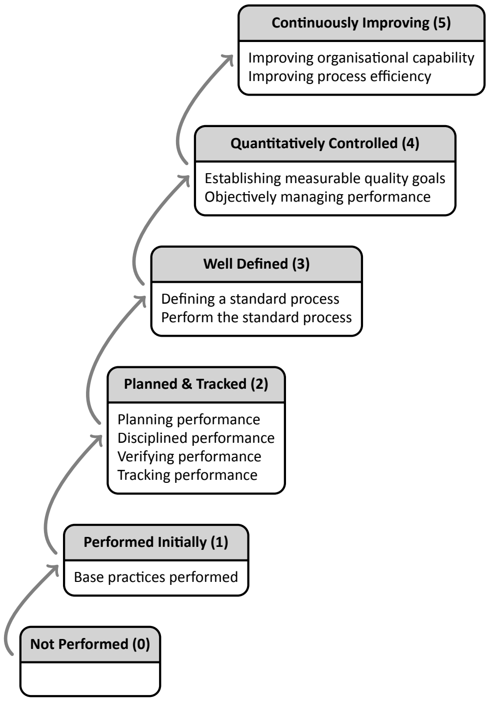
The **Capability Maturity Model** (Fig. 4) was originally developed by Carnegie Mellon as a tool to help choose between software companies when putting a development contract out to tender. Companies with comprehensive systems in place to optimize their operations are said to be most mature, while those with more ad hoc methods are said to be immature. The version shown here has been specialised for systems engineering, but ANDS has produced a version relating to institution-level research data management, while Crowston and Qin from Syracuse University have produced a version for project-level scientific data management.

The **Three-Legged Stool** (Fig. 5) is a concept that came out of Cornell for measuring institutional readiness for digital preservation. It was developed into a balanced scorecard measure, known as AIDA, and later adapted for research data management in the DCC's CARDIO tool. It's the CARDIO version shown here.

One last example is the **Software Maturity Curve** (Fig. 6), which shows the stages through which species of software applications naturally progress over time. The shape resembles the well known hype curve but this one refers to the proliferation or otherwise of software tools that all do approximately the same job.

3 Community Capability Model Framework (CCMF)

So now we come to the framework that we proposed in the White Paper. It was developed through consultation with various communities, by means of:


Continuously Improving (5)
Improving organisational capability
Improving process efficiency
Quantitatively Controlled (4)
Establishing measurable quality goals
Objectively managing performance
Well Defined (3)
Defining a standard process
Perform the standard process
Planned & Tracked (2)
Planning performance
Disciplined performance
Verifying performance
Tracking performance
Performed Initially (1)
Base practices performed
Not Performed (0)

- The enterprise is divided into *process areas* (e.g. Ensure Quality, Manage Risk).
- Achieving a *capability level* within a process area means implementing a certain set of practices.
- These practices are grouped into *common features* (see figure).
- At Level 1, each process area has its own set of *base practices*.
- At Levels 2--5, all process areas share sets of *generic practices*.

http://www.sei.cmu.edu/reports/95mm003.pdf

Figure 4: Systems Engineering Capability Maturity Model

Organisation

1. Data Ownership and Management
2. Data Policies and Procedures
3. Data Policy Review
4. Sharing of Research Data/Access to Research Data
5. Preservation and Continuity of Research
6. Internal Audit of Research Activities
7. Monitoring and Feedback of Publication
8. Metadata Management
9. Legal Compliance
10. Intellectual Property Rights and Rights Management
11. Disaster Planning and Continuity of Research

Technology

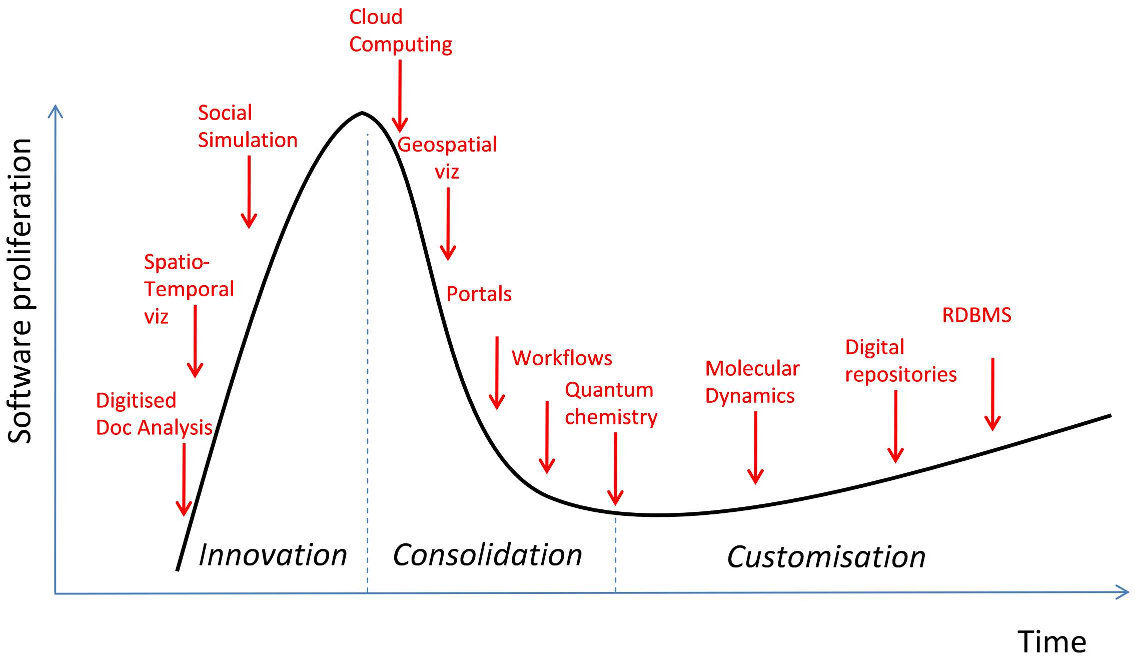
1. Technological Infrastructure
2. Appropriate Technologies
3. Ensuring Availability
4. Managing data integrity
5. Obsolescence
6. Managing technological change
7. Security Provisions
8. Security Processes
9. Metadata tools
10. Institutional Repository

Resources

1. Data Management Costs and Sustainability
2. Business Planning
3. Technological Resources Allocation
4. Risk Management
5. Transparency of Resource Allocation
6. Sustainability of Funding for Data Management and Preservation
7. Data Management Skills
8. Number of Staff for Data Management
9. Staff Development Opportunities

http://cardio.dcc.ac.uk/

Figure 5: The Three Legged Stool



https://ocs.arcs.org.au/index.php/eraust/2010/paper/view/141

Figure 6: Software Maturity Curve

Case studies

- A UK funder: the Economic and Social Research Council
- P.I.s and research leaders from eResearch South Consortium
- University of Bath (Pro-VC for Research, Computing Services, Research Office,...)

Workshops

- York
- Harvard
- Bristol
- Stockholm
- Melbourne

The framework (Fig. 7) consists of eight factors that cover the human, technical and environmental characteristics of a community. Within each factor, we've identified some quite specific characteristics which we think are important and relevant for community capability, so I'll show them to you now. I'll start with Collaboration and work round clockwise.

3.1 Collaboration

One of the things we'll come back to again and again is that data-driven research is resource intensive: you need a lot of data, preferably mashed up from multiple sources, and a lot of processing power. The measures we pick, for the most part, come back to that. So, in communities where there's a lot of collaboration you are more likely to get data and computing resources at a scale where data-driven research can flourish. We've identified several different types of collaboration that might be important.

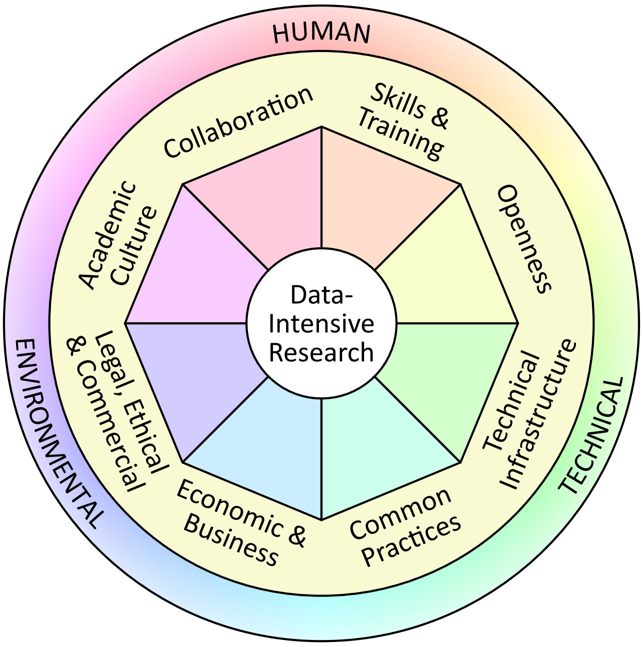


Figure 7: The Community Capability Model Framework

Collaboration within the discipline/sector

Lone researchers.	Departmental research groups.	Collaboration across research groups within or between organisations.	Discipline organised at a national level.	International collaboration and consortia.

The first is collaboration within a discipline, and the more the better, really, though you'd measure this as a curve plotted against research stage rather than as a single value.

Collaboration and interaction across disciplines

No collaboration with other disciplines.	Individual researchers occasionally collaborate outside their discipline.	Disciplines collaborate through joint conferences or publications.	Bilateral collaborations.	Formal collaboration between research groups from several different disciplines.

Then we have inter- and mutli-discplinary research. Again, the more the better as you're more likely to get new insights if you combine data that have never been combined before.

Collaboration and interaction across sectors

None.	Attempts have been made but are not considered successful.	Despite successful examples working with other sectors is not the norm – some barriers are perceived.	A discipline or group has gained experience of working closely with one or two sectors.	Work successfully with several other sectors on different problems.

The same applies to cross-sectoral collaboration, where more also implies better or more sustained resourcing.

Collaboration with the public

No collaboration with the public.	The public's involvement is limited to acting as subjects of study, user testing, etc.	Contact with the public is only through occasional appearance in the media e.g. news bulletins, TV programmes.	Mainly informational, sometimes participative, targeted media programmes are organised to engage the public e.g. science fairs.	Dedicated programmes involving the public in research; crowd sourcing/citizen science.

Lastly, the public can provide data or processing resources on a scale you wouldn't otherwise be able to achieve. I'm thinking of wildlife surveys or projects like Galaxy Zoo where a huge team of volunteers did the job of pattern recognition software.

3.2 Skills and training

Here we're looking at the skills researchers need in order to perform data-driven research.

Skill sets

Tools and technologies (cloud computing, visualisations, statistical analysis, simulations, modelling).	Data description and identification (metadata, vocabularies, citation).	Collaboration and communication (engaging with other researchers, the public, the media).	Policy and planning (data management, business models).

We've identified four skill sets that we think are important.

Pervasion of training

No training available.	Training programmes in development.	Training available but not embedded within u/g and p/g degree programmes. Patchy uptake. Little or no on-job coaching or mentoring on data management.	Training embedded within u/g and p/g degree programmes and available for researchers. Mentors usually provided on request.	Dedicated training, fully embedded in all u/g and p/g degree programmes, accredited with professional qualifications, and an established part of continuing professional development.

And then we look at how well these skills pervade the research community; the more pervasive the better.

3.3 Openness

With openness we come back to the idea that data-driven research works best when there is a pool of open data that researchers can draw from. But it's not just quantity that's the issue, it is also things like quality, interoperability and documentation.

Openness in the course of research

No sharing. No details released.	Selected details released, e.g. in a proposal or project plan.	Selected intermediate results are shared within a limited group.	Intermediate results are shared through traditional means, e.g. conference papers.	Sharing is done publicly on the web. Full details are disclosed.

We think researchers are more likely to get all that right if they open up their data early, because if problems come to light after they've moved on to other things it's too late by then.

Openness of published literature

No sharing of papers or metadata outside publication channels.	Authors share metadata for their publications.	Authors share theses or other selected sections from the literature.	Authors provide copies of their publications on request or other negotiated means.	Publications are made available on open access.

The published literature is often the place where researchers advertise their data and document how they collected them, so having those publications openly available is very helpful for driving up reuse.

Openness of data

No sharing. No details released.	The data are described in the literature but not made available.	Data are available on request, after embargo or with other conditions.	Efforts are made to make data discoverable and re-usable as well as available.	Data is available in re-usable form and freely available to all. Community curation of the data may be possible.

This should be self-explanatory.

Openness of methodologies/workflows

No sharing. No details released.	Released within limited scope. Partial details released.	The details of the workflow are shared but not the underlying scripts; only partial stages of the workflow are shared.	Sharing publicly on the web. Non-standard scripts, tools and software released.

Sometimes to expand a dataset you need to process some new raw data according to an existing workflow. In such cases the existing workflow has to be openly available.

Reuse of existing data

Only own data used.	Data exchanged within limited scope.	Regularly combine data sets in specific established ways. Provenance tracked in ad hoc ways.	Multiple existing datasets often combined. Provenance tracked systematically.

Finally, it's one thing making all these things open, but it doesn't do any good until people start using them, so we thought it important to measure that.

3.4 Technical infrastructure

Moving on, we come to tools and platforms. It's an obvious point but it needs making: if a community is to perform data-driven research, it needs a technical infrastructure that is up to the job. So we identified the major tasks for which tools are needed, and measures of their adequacy for current and future research.

Computational tools and algorithms

None.	Tools exist but perform below requirements.	Tools have sufficient features to meet the needs of most users.	Tools have features few people use, expected to meet users' needs for the next few years.

These are probably the most fundamental for data-intensive research.

Tool support for data capture and processing

No tool support for data capture.	Tools do not meet user requirements well or do not interoperate. Tools are custom and quality varies.	One or two good tools available. A few clear leaders.	Most tools that support data capture do it well and meet user requirements.	All tools support data capture well and interoperate. There is a good choice of tools for data processing.

With capture tools the major issue is interoperability. They really need to spit data out in a format that can be widely used.

Data storage

None.	Insufficient data storage available to meet user needs.	Although data storage capacity is sufficient, other requirements (e.g. security) are not met.	Dedicated storage facilities meet current requirements, but will be outgrown shortly.	Storage is available and is expected to meet future needs.

This isn't just about quantities but also reliability, security and so on.

Support for curation and preservation

None.	Support is only available in specialised cases.	Insufficient tools and facilities exist to meet needs.	Dedicated tools are available and are widely used.	Common infrastructure is well funded and well used.

These are tools for keeping data as useful as possible both now and in the long term.

Data discovery and access

None.	Discovery services very discipline-specific; require specialised knowledge or rights.	Discovery opened to all but siloed (not interopeable).	Data discoverable and accessible to all, good integrated services.

It's no good having data if it can't be found, and this is a particularly tricky problem to tackle in the context of interdisciplinary research.

Integration and collaboration platforms

None.	Platforms exist but perform below requirements.	Platforms have sufficient features to meet the needs of most users.	Platforms have features few people use, expected to meet users' needs for the next few years.

The better these are, the more capacity researchers have for data-intensive research.

Visualisations and representations

None.	Tools exist but perform below requirements.	Tools have sufficient features to meet the needs of most users.	Tools have features few people use, expected to meet users' needs for the next few years.

As the Fourth Paradigm is all about spotting unexpected patterns in data, it's clear that good visualisation tools are important.

Platforms for citizen science

None.	Customised tools available, used by a small number of groups.	Very flexible tools available and well used.	Tools have been re-deployed to other disciplines.

And as a companion to what I said earlier about citizen science, you can only get the benefit of mass lay participation if you have a platform for managing it.

3.5 Common practices

By Common Practices we mean standards and conventions that make it easier to share, use and recombine data. It’s not just about the quantity of such practices, as too many can be just as divisive as too few, so we’ve worded our measures carefully. Again, we’ve split this factor up into tasks.

Data formats

No standard formats available: ad hoc formats proliferate.	Standard formats are in development but not yet in use.	Some standard formats available but not widely adopted or community begins to converge on small number of formats.	Standard formats are widely adopted for some but not all types of data.	Standard formats are universally adopted for all types of data. Faithful conversions are possible between ‘rival’ standards.

I don’t think I need to say any more about data formats.

Data collection methods

Methods are not usually shared.	Methods are shared but not widely reused.	Agreed methods are in development.	Although some methods are agreed there are gaps in the methods covered or room for improvement in the quality.	Methods are well known, well documented and well used.

We’re including everything from how performances are recorded to how sensors are used.

Processing workflows

Workflows are not usually shared.	Workflows are shared but not widely reused.	Agreed workflows are in development, or community begins to converge on a small number of workflows.	Agreed workflows are available with some gaps, or room for improvement in quality.	Several standardised workflows widely used.

Again, if data have all been processed in the same way they can be combined more easily.

Data packaging and transfer protocols

Packaging and transfer performed ad hoc.	Standard protocols are in development but not yet in use.	Some standard protocols available but not widely adopted or community begins to converge on small number of protocols.	Some standard protocols available with some gaps, or room for improvement in quality.	One or two standardised formats/protocols widely used.

This is about getting the data from one user to another.

Data description

No standard metadata schemes exist.	Standard metadata schemes are in development but not yet in use.	Some metadata schemes are published and recognised, but with little uptake or known flaws.	Recognised metadata schemes agreed, with some gaps.	Mature, agreed and widely used metadata schemes exist.

This is about the community agreeing what it needs to know about data in order to reuse it.

Vocabularies, semantics, ontologies

No standard schemes are available.	Some schemes are published but they are experimental with limited uptake.	Standards are being actively developed; agreement and standardisation by the community is being pursued.	Some standard schemes are available, however gaps still exist.	Standard schemes are mature with good take-up by the community and widely applied.

These are other things that can affect interoperability.

Data identifiers

None in use. Some used experimentally. Sporadic use.	Some trustworthy identifiers adopted.	Discipline-specific identifiers widely used.	International, well managed, sustainable schemes routinely used.

Data identifiers help with retrieval, access, and fulfilling legal attribution requirements, of which more later.

Stable, documented APIs

APIs not generally published or used.	Some tools offer APIs but with insufficient documentation.	A handful of well recognised APIs but these are the exception rather than the norm.	Most key disciplinary tools and services have useful, stable, and documented APIs.	Culture of developing APIs widespread.

Stable APIs are particularly useful in the sorts of automated workflows that characterise some forms of data-driven research.

3.6 Economic and business models

We recognise that data-driven research can be quite expensive due to the data volumes and processing power involved, so it is important that the funding is in place to support it. In the first six measures, we look at the sustainability, geographic scale and size of funding for research…

Sustainability of funding for research

Funding focused on short-term projects and quick returns.	Single-phase thematic investments on a 3-5 year timescale.	Multi-phase thematic investments in 5-10 year blocks which build a community.

Geographic scale of funding for research

Projects funded internally or through grants from regional agencies.	Projects funded by national funders.	Funding by international bodies and bi-lateral initiatives between national funders.

Size of funding for research

Small-scale projects (e.g. to exploit open innovation methodologies for bio-informatics tool development).	Mid-scale projects (e.g. digitisation and analysis of large textual corpora).	Major investment (e.g. in longitudinal data surveys).

…and for infrastructure. Note that alignment between these different measures can be just as important as the absolute values: if the funding is distributed but the facilities need to be central, that can cause problems.

Sustainability of funding for infrastructure

One-off investments with no commitment to sustainment.	Infrastructure projects allowed slow transition to self-financing model.	Sustained multi-decade investments in data centres and services.

Geographic scale of funding for infrastructure

Investments by a single funding body at regional or national level.	Collaborative development at the national level by multiple funders.	Collaborative development between international funders.

Size of funding for infrastructure

Small-scale tool development.	Co-ordinated investments in distributed systems.	Large central investments in network infrastructure or tools.

Public–private partnerships

None.	Informal collaboration with industry but no funding involved.	Corporate/SME are non-funded partners in proposals with academia.	Research is co-funded by industry and other sources.	Established formal co-investment partnerships running long-term multi-phase projects.

Private investment may help with the sustainability of facilities but be detrimental to data sharing.

Productivity and return on investment

Long lead times between project start and submission of outputs (e.g. 6 years), and between acceptance and publication of papers (e.g. 2 years). Funders expect projects to publish a small number of papers each with high long-term impact.	Mid-range lead times between project start and submission of outputs (e.g. 3 years), and between acceptance and publication of papers (e.g. 1 year). Funders expect projects to publish a moderate number of papers in high impact journals.	Short lead times between project start and submission of outputs (e.g. 18 months), and between acceptance and publication of papers (e.g. 3 months). Funders expect projects to publish a large number of both high quality papers and progress reports.

This last factor is about the character of research and has a complex relationship to capability. All I'll say for now is that when the research lifecycle is longer, we should expect data-driven research to take longer as well.

3.7 Legal and ethical issues

We thought it important to draw out legal and ethical issues because, when they get in the way of data sharing and reuse, they are a lot harder to surmount than technical barriers.

Legal and regulatory frameworks

No coordinated response to legal, regulatory and policy issues. Confusion over obligations is widespread.	Basic frameworks exist but they are disjointed and frequently more hindrance than help.	Moderately sophisticated and helpful frameworks exist, but awareness of them is poor and the corresponding procedures are not well enforced.	Robust frameworks and procedures exist and are regulated at institutional level, but researchers do not fully trust them.	Trusted frameworks and procedures are in place. Discipline is well regulated by disciplinary bodies, professional societies.

What we intend by legal and regulatory frameworks are systems for guaranteeing, effectively, the legality of data sharing and reuse. Institutions are rightly risk-averse, so removing those risks benefits everyone.

Management of ethical responsibilities and norms

No standard procedures in place. Poor or uneven awareness of ethical issues and how to approach them.	Some procedures exist but they lack consistency, may hinder rather than help, and are rarely followed.	Consistent and useful procedures exist but they are not enforced.	Robust procedures are in place and are enforced locally, though they may be seen as a burden.	Trusted and accepted procedures are in place, and are enforced at the national or international level.

The second measure is the equivalent but for ethical matters.

3.8 Academic culture

Our last factor is a catchall, really, for aspects of academic culture that have not already been covered.

Entrepreneurship, innovation and risk

Highly risk-averse.	Moderately risk averse.	Calculated risks taken.	Moderately innovative and experimental.	Highly innovative and experimental.

Moving to a new paradigm is inherently risky, as it involves investments that may or may not pay off, so the leap will be easier where such risks are tolerated.

Reward models for researchers

None.	Narrow range of contributions recognised.	Wider range of contributions recognised, but informally.	Measures exist for more than one type of contribution and are well recognised.	All contributions are recognised and rewarded, through established procedures and measures.

Under reward models, the main things we're looking for are incentives for researchers to share their data and reuse other people's.

Quality and validation frameworks

Lightweight self-review of data. Results not reproducible.	Lightweight review of data by colleagues.	Partial peer review (whether data matches description, whether column headings make sense).	Thorough peer review (for integrity, appropriateness, reproducibility).	Data thoroughly reviewed and curated by specialists. Results are reproducible.

And the last measure reiterates that even if data are shared, they won't get reused if they are of low quality.

If all that seems a lot to take in, don't worry, we thought so too. So since publishing the White Paper we have been working on a tool that adapts the framework for particular stakeholders. For more on that I shall hand you over to my colleague Manjula.

4 Credits

Project Team

UKOLN

Liz Lyon
Alex Ball
Michael Day
Monica Duke
Manjula Patel
Michelle Smith

Microsoft Research

Kenji Takeda
Alex Wade

Website: `http://communitymodel.sharepoint.com/`

Acknowledgements

The CCMDIR Project is funded by Microsoft Research.

UKOLN is funded by JISC, as well as by project funding from JISC, Microsoft Research and the European Union. UKOLN also receives support from the University of Bath where it is based.