Introduction to the Community Capability Model Framework & Profile Tool

Community Capability Profiling Workshop
Introducing a new tool to facilitate Data-Intensive Research

International Digital Curation Conference 2014
Monday 24th February 2014
San Francisco, US

Liz Lyon, iSchool, University of Pittsburgh, US
Manjula Patel, UKOLN Informatics, University of Bath, UK

Unless otherwise stated this work is licensed under a
Creative Commons Attribution-ShareAlike 3.0 Unported License.

Workshop Goals

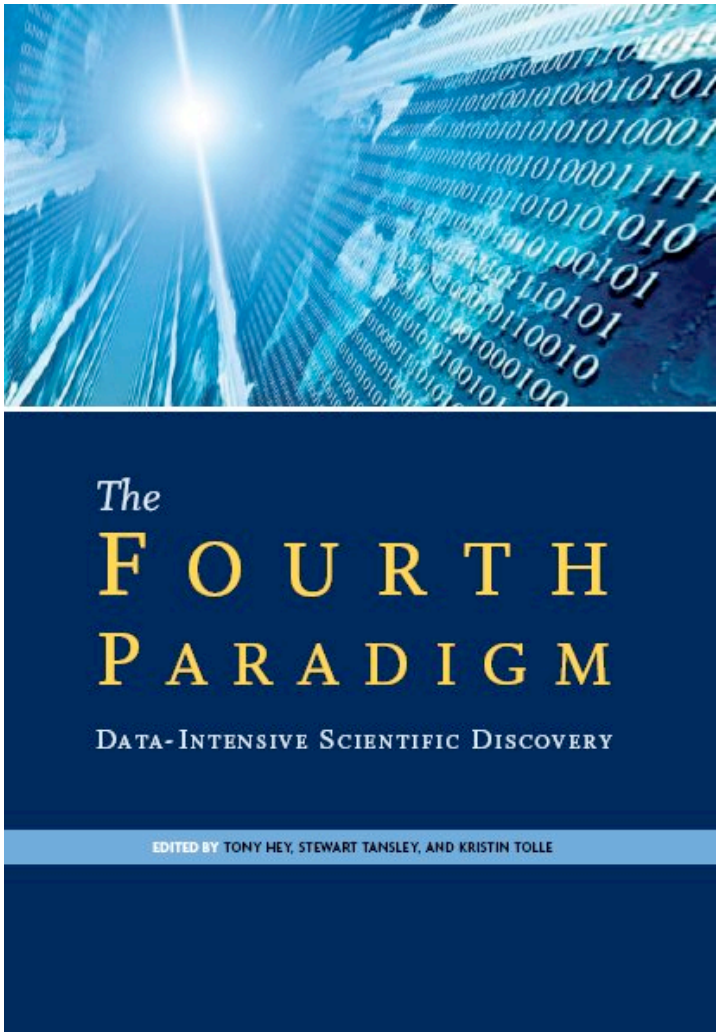
- To demonstrate application of the CCMF Profile tool across a range of contexts and communities: disciplines, organisations, groups etc.
- To collect completed profiles from participants in a diverse range of disciplines and sub-disciplines
- To investigate opportunities to customise the Profile for particular domains

Context

- Experimental Science
  - Observational description of natural phenomena
- Theoretical Science
  - Use of models and equations

    e.g. Newton's Laws
- Computational Science
  - Digital simulation of complex phenomena
- Data-Intensive Science
  - Unify experiment, theory and simulation

    - Jim Gray
- Data-Intensive Research
  - Intensive data collection and processing
  - Large quantities of data
  - Combination of individual datasets



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

Motivations for DIR

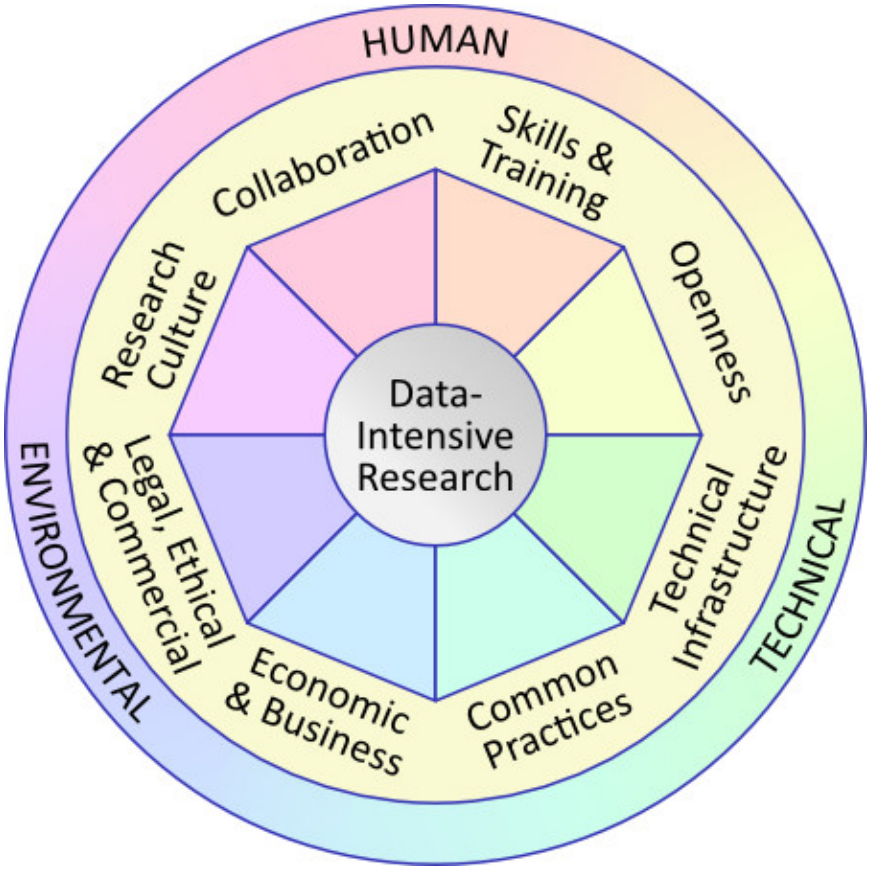
- Funding Bodies (e.g. NSF, European Union, UK Research Councils, Trusts, Learned Societies, Companies, Foundations)
  - Derive maximum research, economic and social benefits from investments
  - Improve the quality and efficiency of research (robust and reproducible)
  - Increase knowledge transfer within discipline; across disciplines; between sectors
  - Build sub-disciplinary, disciplinary and inter-disciplinary communities
  - Develop added-value services based on corpora of research data
- Institutions (e.g. HEIs, Facilities (e.g. CERN, STFC, EMBL))
  - Improve the quality and efficiency of research (robust and reproducible)
  - Increase ability to attract research funds
  - Build institutional and cross-institutional communities
  - Develop added-value services based on corpora of research data
  - Include data citation into research evaluation systems e.g. UK's REF
- Researchers (Principal Investigators)
  - Opportunities for new and innovative research
  - Improve the quality of research (robust and reproducible)
  - Improve citations and reputation
  - Career advancement
  - Add data citation into research evaluation systems e.g. UK's REF

Data-Intensive Research Lifecycle

Areas that need particular attention

- Legal, ethical and commercial issues
  - IPR, privacy, sensitivity, licensing
- Gaining informed consent for reuse and repurposing
- Appraisal and quality control
  - Collection and acquisition policies, peer review
- Trustworthiness
  - Metadata, documentation, context, provenance
- Scale and complexity of data
  - Workflows, methodologies, software, OAIS Representation Information
- Publication and sharing
  - Release policy, controlled access, indexing, interoperability (syntax and semantics), cross-searching, federation
- Citation, attribution and accreditation in scholarly communications
  - granularity, versioning, persistent identifiers

The CCMF



communitymodel.sharepoint.com

- The Community Capability Model Framework (CCMF)
  - Profiling current readiness or capability of a community for DIR
  - Indicating priority areas for change and investment
  - Developing roadmaps for achieving a target state of readiness

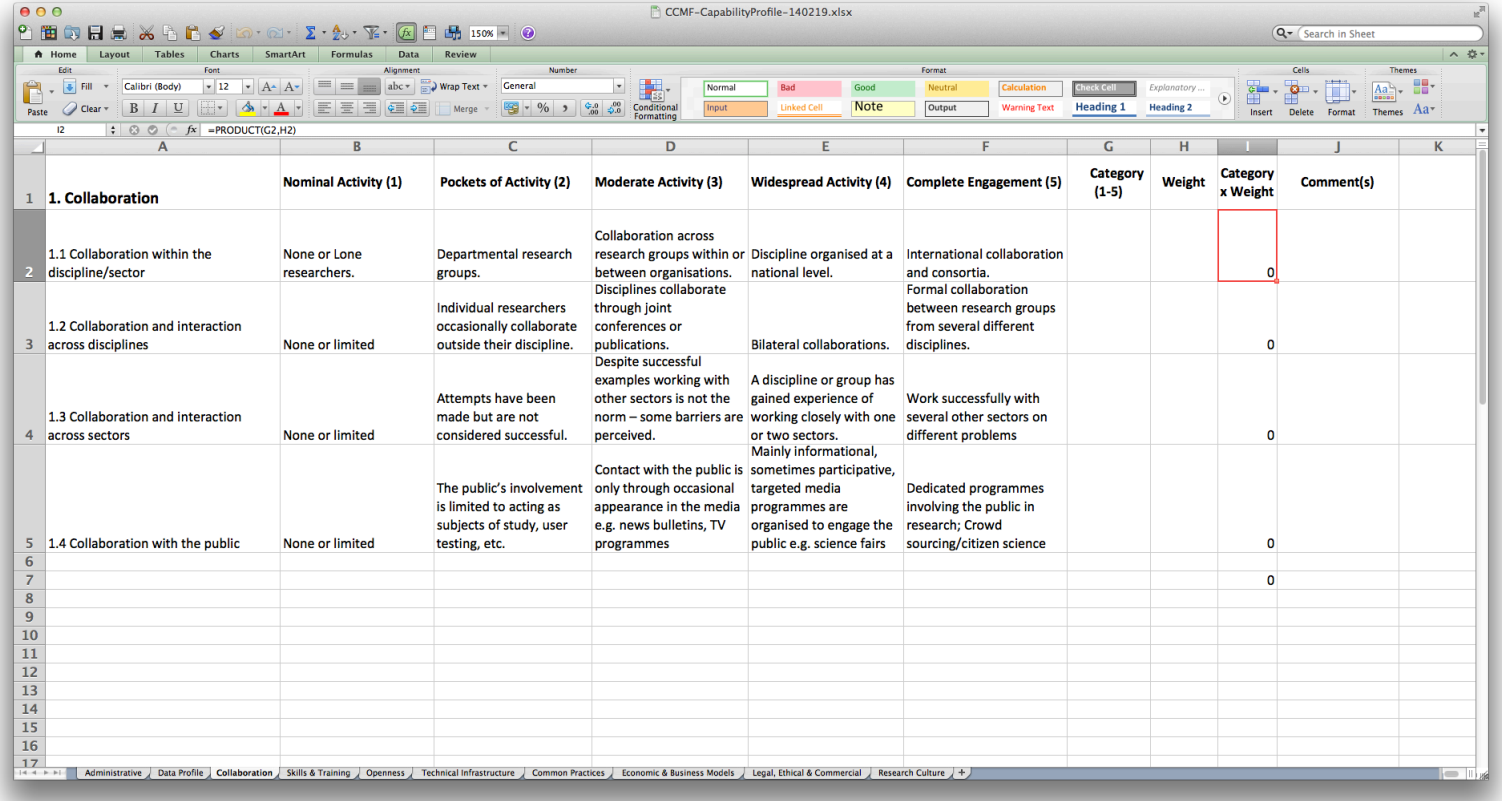
  CCMF White Paper, April 2012
- Developed through consultation: case studies and workshops
- Primarily a tool for self-assessment
- Categorised into Environmental, Human and Technical elements with eight factors:

Openness	Legal, Ethical & Commercial
Collaboration	Economic & Business
Skills & Training	Common Practices
Research Culture	Technical Infrastructure

- Each factor has *characteristics* associated with it

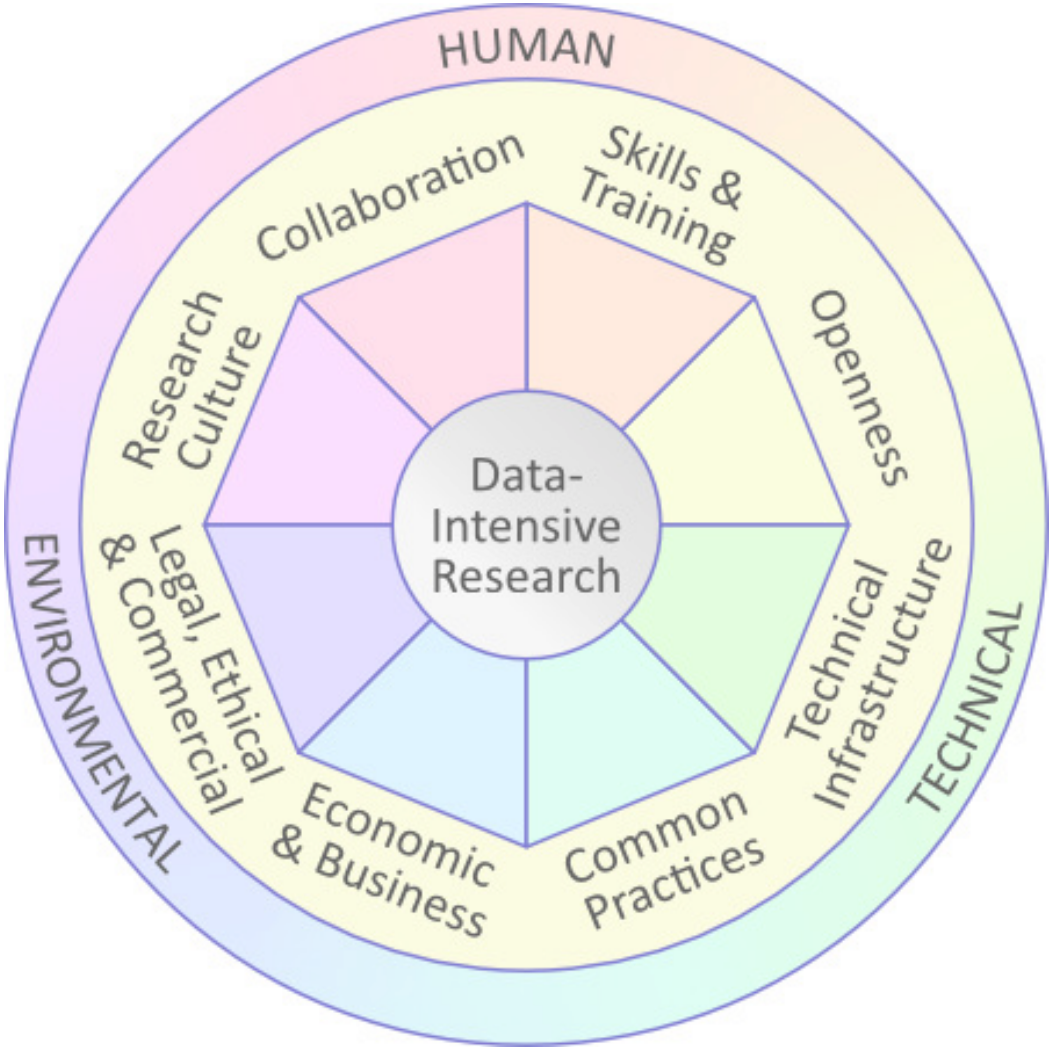
CCMF Profile Tool

- Based on CCMF
- Implemented as an MS Excel spread sheet
- Separate worksheets for each of the eight CCMF factors
- A scorecard tool (5 levels or *dimensions* for each *characteristic* within each *factor*)



	A	B	C	D	E	F	G	H	I	J	K
1	**1. Collaboration**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**	
2	1.1 Collaboration within the discipline/sector	None or Lone researchers.	Departmental research groups.	Collaboration across research groups within or between organisations.	Discipline organised at a national level.	International collaboration and consortia.			0		
3	1.2 Collaboration and interaction across disciplines	None or limited	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.			0		
4	1.3 Collaboration and interaction across sectors	None or limited	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			0		
5	1.4 Collaboration with the public	None or limited	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			0		
6											
7									0		
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											

CCMF Profile Tool Worksheets



Group Work

- Work with colleagues in similar or associated domains
- Complete the Profile for your chosen domain
- Review the Profile and make recommendations for enhancements in your domain e.g. semantics, exemplars
- Discuss results in feedback session
- Download CCMF Profile tool:

  people.bath.ac.uk/lismp/CCMF/CCMF-Profile.xlsx

Acknowledgements

This work is funded by Microsoft Research Connections.

UKOLN Informatics receives additional support from the University of Bath where it is based.

Contacts:

Liz Lyon: elyon@pitt.edu

Kenji Takeda: kenjitak@microsoft.com

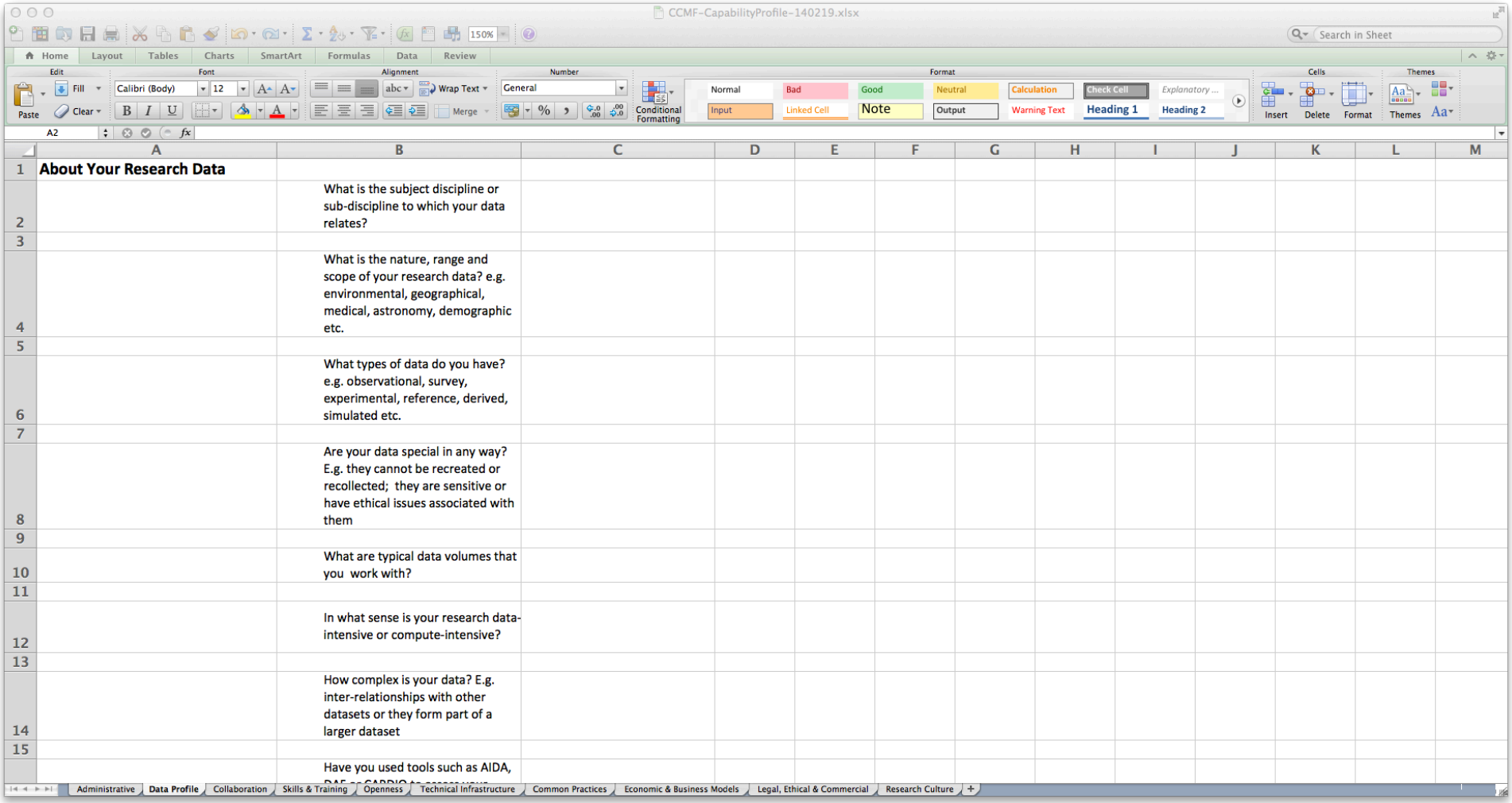
Manjula Patel: m.patel@ukoln.ac.uk

Further Information:

http://communitymodel.sharepoint.com/

Additional Slides

Data Profile

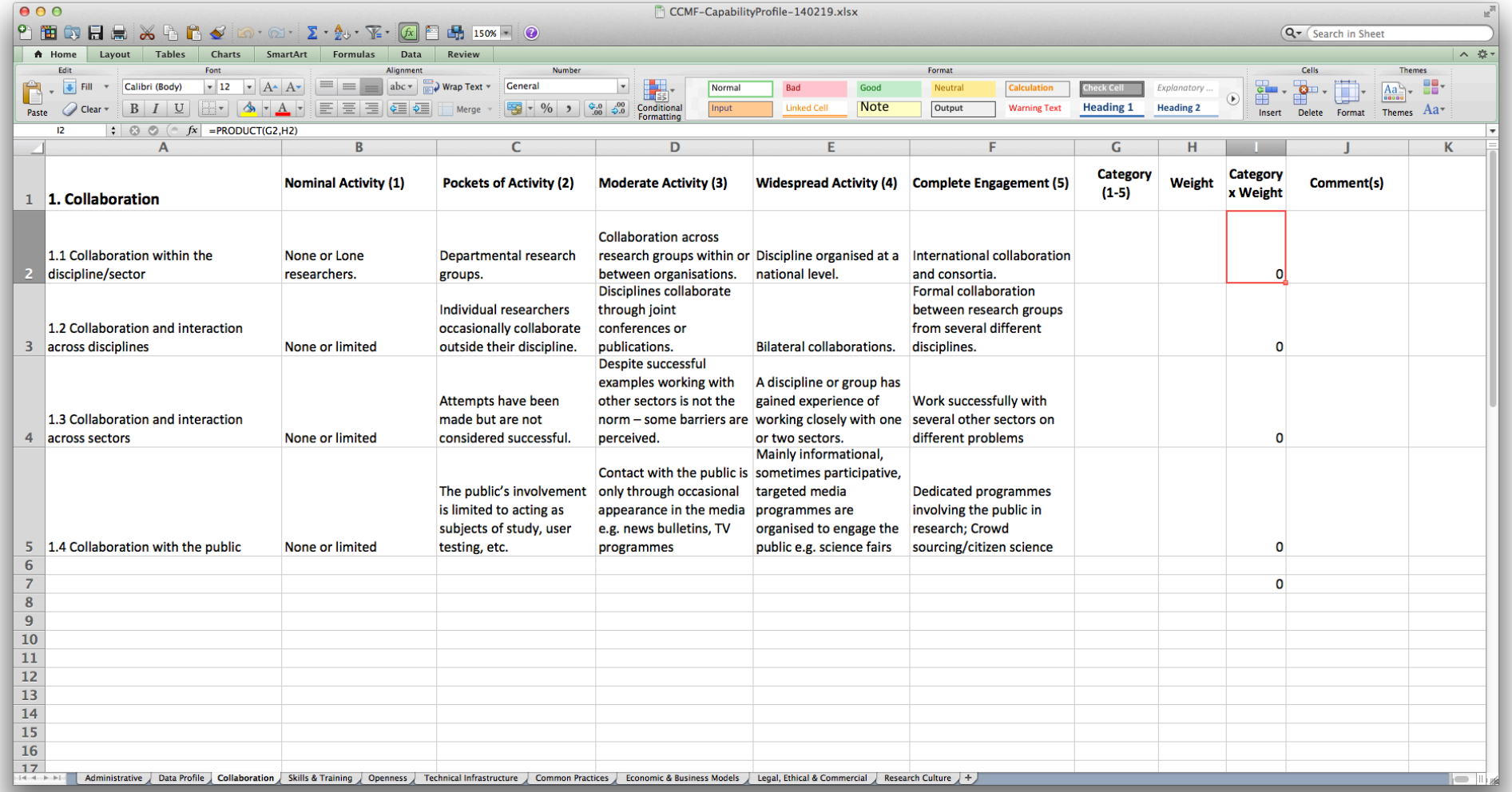


CCMF-CapabilityProfile-140219.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**About Your Research Data**												
2		What is the subject discipline or sub-discipline to which your data relates?											
3													
4		What is the nature, range and scope of your research data? e.g. environmental, geographical, medical, astronomy, demographic etc.											
5													
6		What types of data do you have? e.g. observational, survey, experimental, reference, derived, simulated etc.											
7													
8		Are your data special in any way? E.g. they cannot be recreated or recollected; they are sensitive or have ethical issues associated with them											
9													
10		What are typical data volumes that you work with?											
11													
12		In what sense is your research data-intensive or compute-intensive?											
13													
14		How complex is your data? E.g. inter-relationships with other datasets or they form part of a larger dataset											
15													
		Have you used tools such as AIDA,											

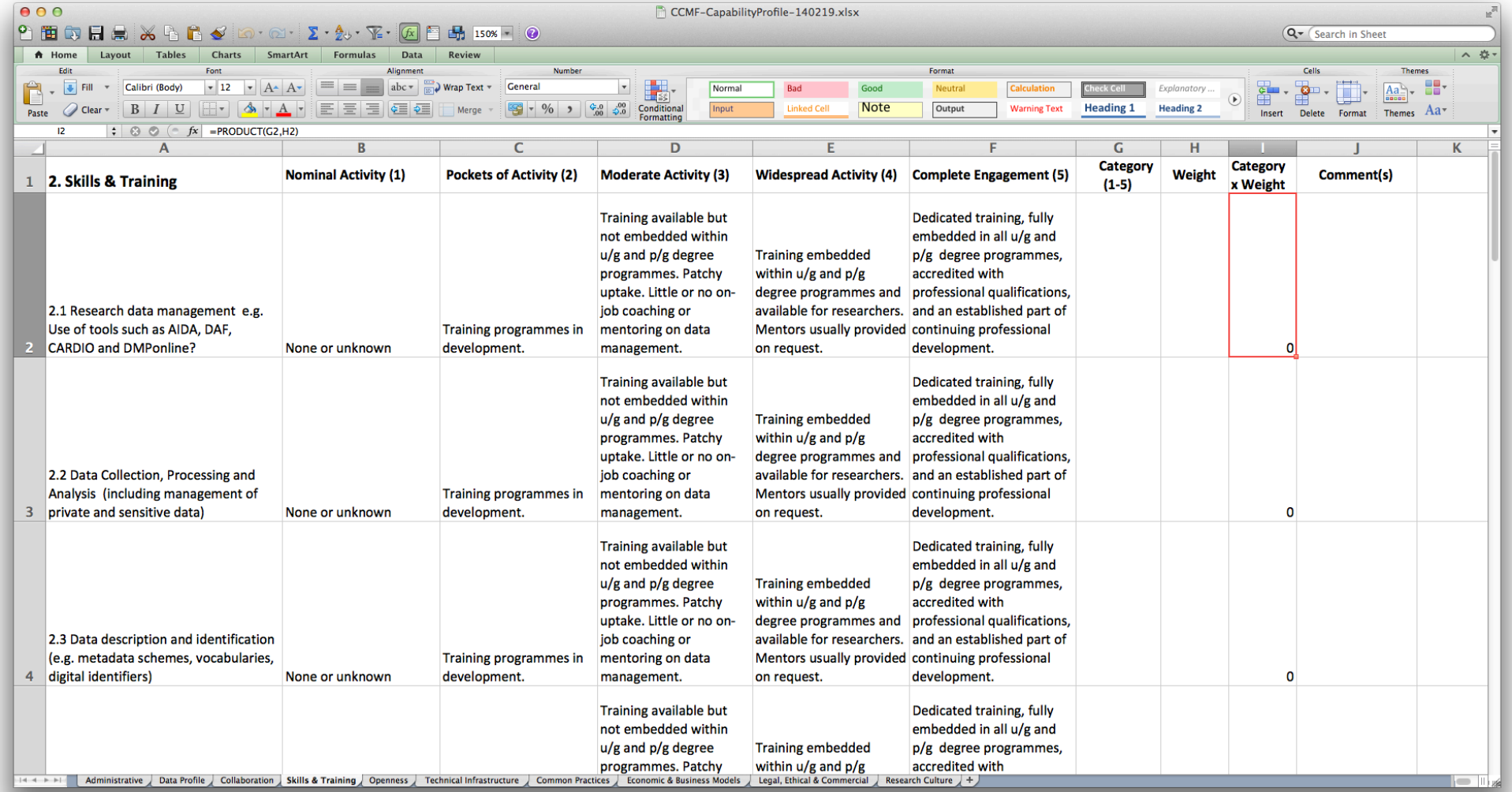
Administrative | Data Profile | Collaboration | Skills & Training | Openness | Technical Infrastructure | Common Practices | Economic & Business Models | Legal, Ethical & Commercial | Research Culture

Collaboration



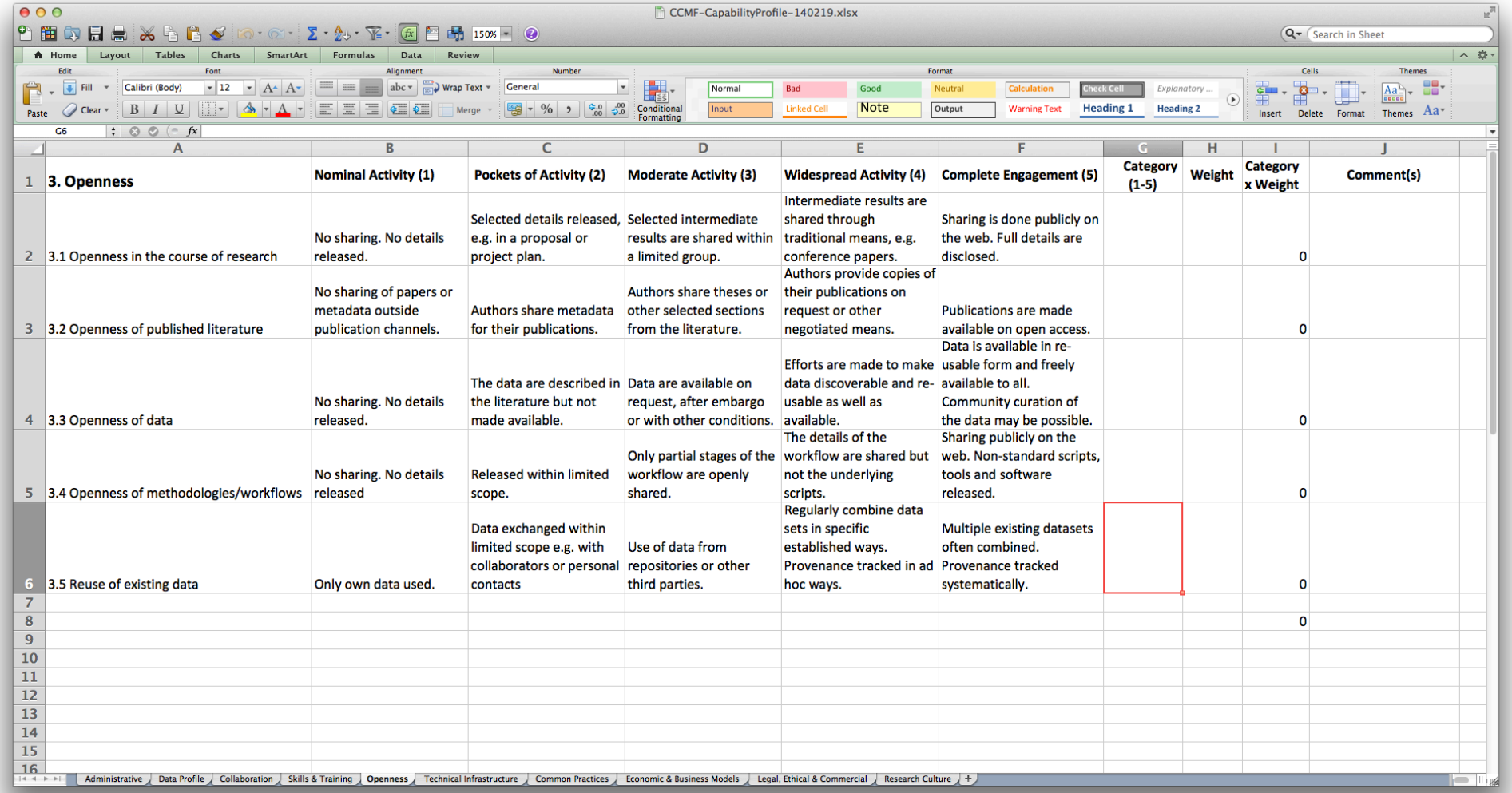
	A	B	C	D	E	F	G	H	I	J	K
1	**1. Collaboration**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**	
2	1.1 Collaboration within the discipline/sector	None or Lone researchers.	Departmental research groups.	Collaboration across research groups within or between organisations.	Discipline organised at a national level.	International collaboration and consortia.			0		
3	1.2 Collaboration and interaction across disciplines	None or limited	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.			0		
4	1.3 Collaboration and interaction across sectors	None or limited	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			0		
5	1.4 Collaboration with the public	None or limited	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			0		
6											
7									0		
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											

Skills & Training



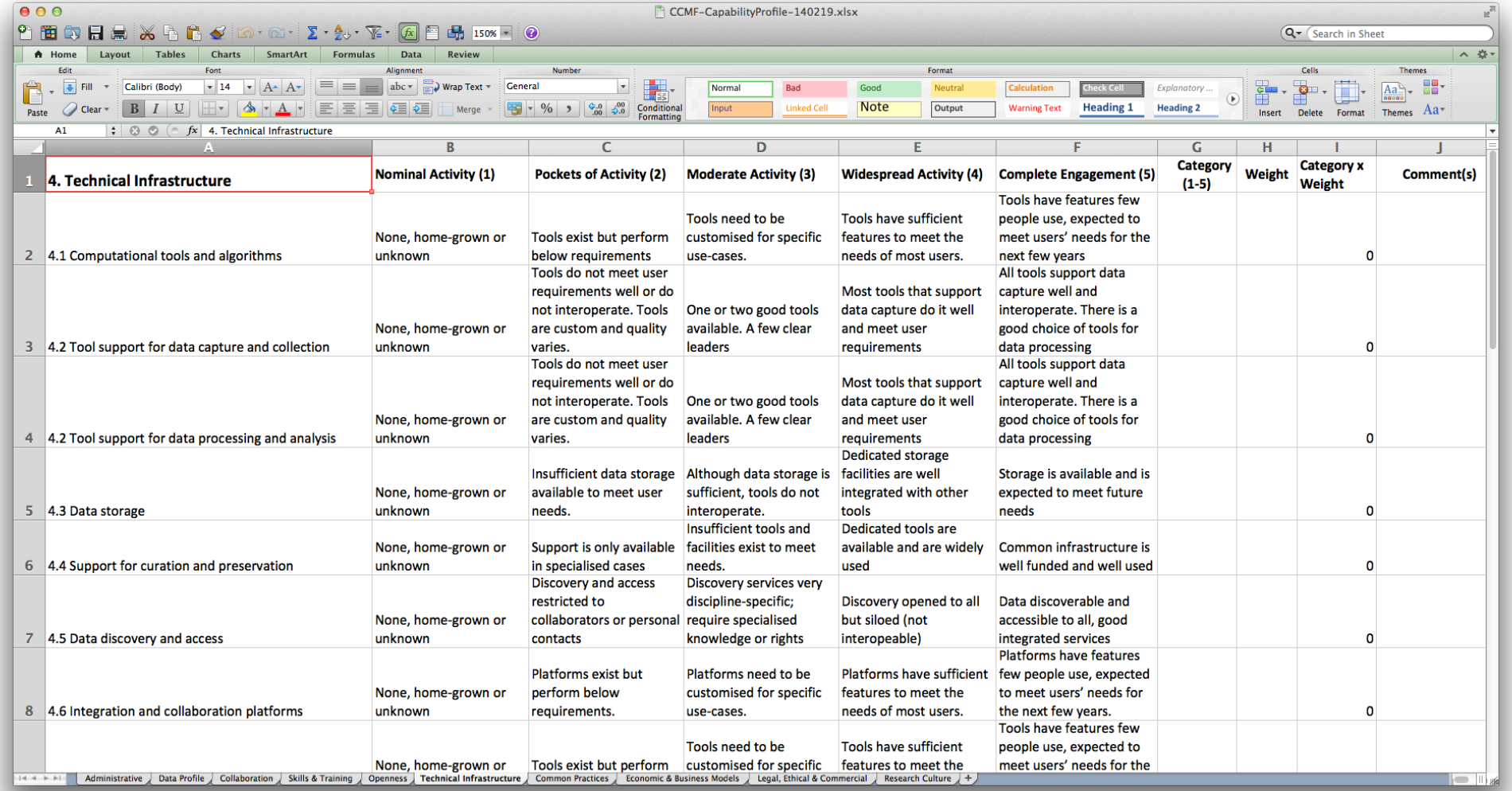
	A	B	C	D	E	F	G	H	I	J	K
1	**2. Skills & Training**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**	
2	2.1 Research data management e.g. Use of tools such as AIDA, DAF, CARDIO and DMPonline?	None or unknown	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.			0		
3	2.2 Data Collection, Processing and Analysis (including management of private and sensitive data)	None or unknown	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.			0		
4	2.3 Data description and identification (e.g. metadata schemes, vocabularies, digital identifiers)	None or unknown	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.			0		
				Training available but not embedded within u/g and p/g degree programmes. Patchy	Training embedded within u/g and p/g	Dedicated training, fully embedded in all u/g and p/g degree programmes, accredited with					

Openness



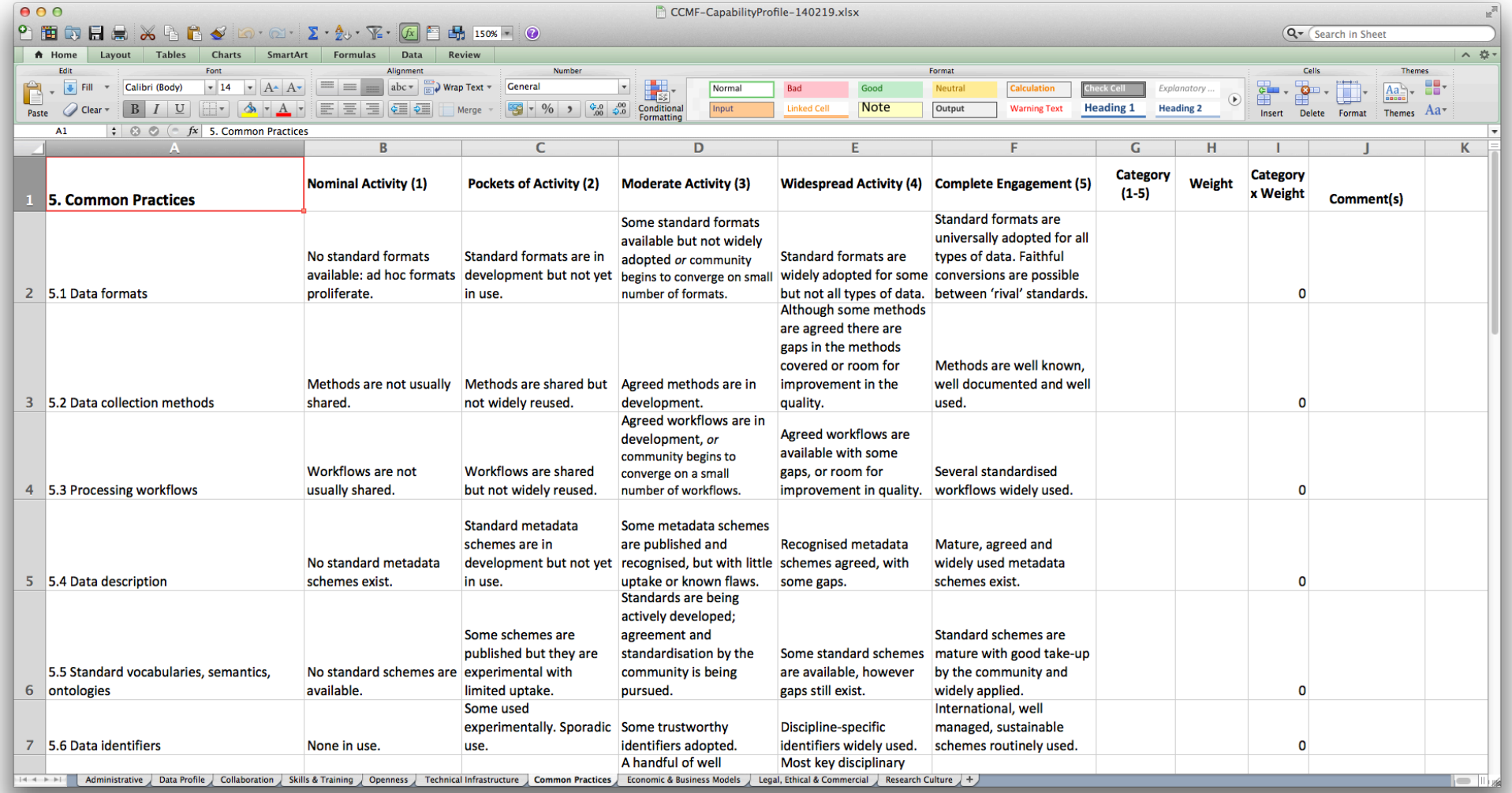
	A	B	C	D	E	F	G	H	I	J
1	**3. Openness**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**
2	3.1 Openness in the course of research	No sharing. No details 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.			0	
3	3.2 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.			0	
4	3.3 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.			0	
5	3.4 Openness of methodologies/workflows	No sharing. No details released	Released within limited scope.	Only partial stages of the workflow are openly shared.	The details of the workflow are shared but not the underlying scripts.	Sharing publicly on the web. Non-standard scripts, tools and software released.			0	
6	3.5 Reuse of existing data	Only own data used.	Data exchanged within limited scope e.g. with collaborators or personal contacts	Use of data from repositories or other third parties.	Regularly combine data sets in specific established ways. Provenance tracked in ad hoc ways.	Multiple existing datasets often combined. Provenance tracked systematically.			0	
7										
8									0	
9										
10										
11										
12										
13										
14										
15										
16										

Technical Infrastructure



	A	B	C	D	E	F	G	H	I	J
1	**4. Technical Infrastructure**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**
2	4.1 Computational tools and algorithms	None, home-grown or unknown	Tools exist but perform below requirements	Tools need to be customised for specific use-cases.	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			0	
3	4.2 Tool support for data capture and collection	None, home-grown or unknown	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			0	
4	4.2 Tool support for data processing and analysis	None, home-grown or unknown	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			0	
5	4.3 Data storage	None, home-grown or unknown	Insufficient data storage available to meet user needs.	Although data storage is sufficient, tools do not interoperate.	Dedicated storage facilities are well integrated with other tools	Storage is available and is expected to meet future needs			0	
6	4.4 Support for curation and preservation	None, home-grown or unknown	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			0	
7	4.5 Data discovery and access	None, home-grown or unknown	Discovery and access restricted to collaborators or personal contacts	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			0	
8	4.6 Integration and collaboration platforms	None, home-grown or unknown	Platforms exist but perform below requirements.	Platforms need to be customised for specific use-cases.	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.			0	
		None, home-grown or	Tools exist but perform	Tools need to be customised for specific	Tools have sufficient features to meet the	Tools have features few people use, expected to meet users' needs for the				

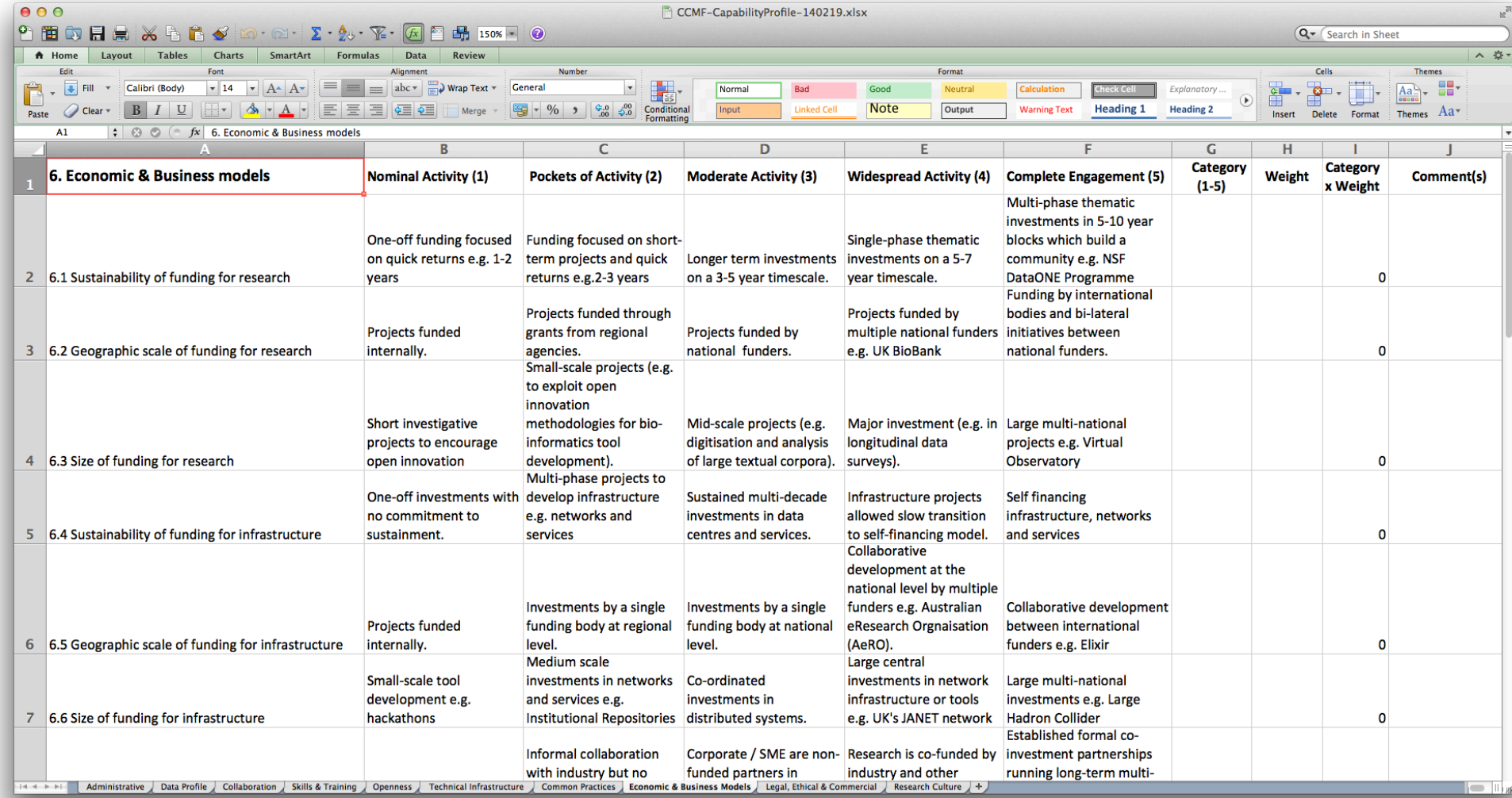
Common Practices


CCMF-CapabilityProfile-140219.xlsx

	A	B	C	D	E	F	G	H	I	J	K
1	**5. Common Practices**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**	
2	5.1 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.			0		
3	5.2 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.			0		
4	5.3 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.			0		
5	5.4 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.			0		
6	5.5 Standard 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.			0		
7	5.6 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.			0		
				A handful of well	Most key disciplinary						

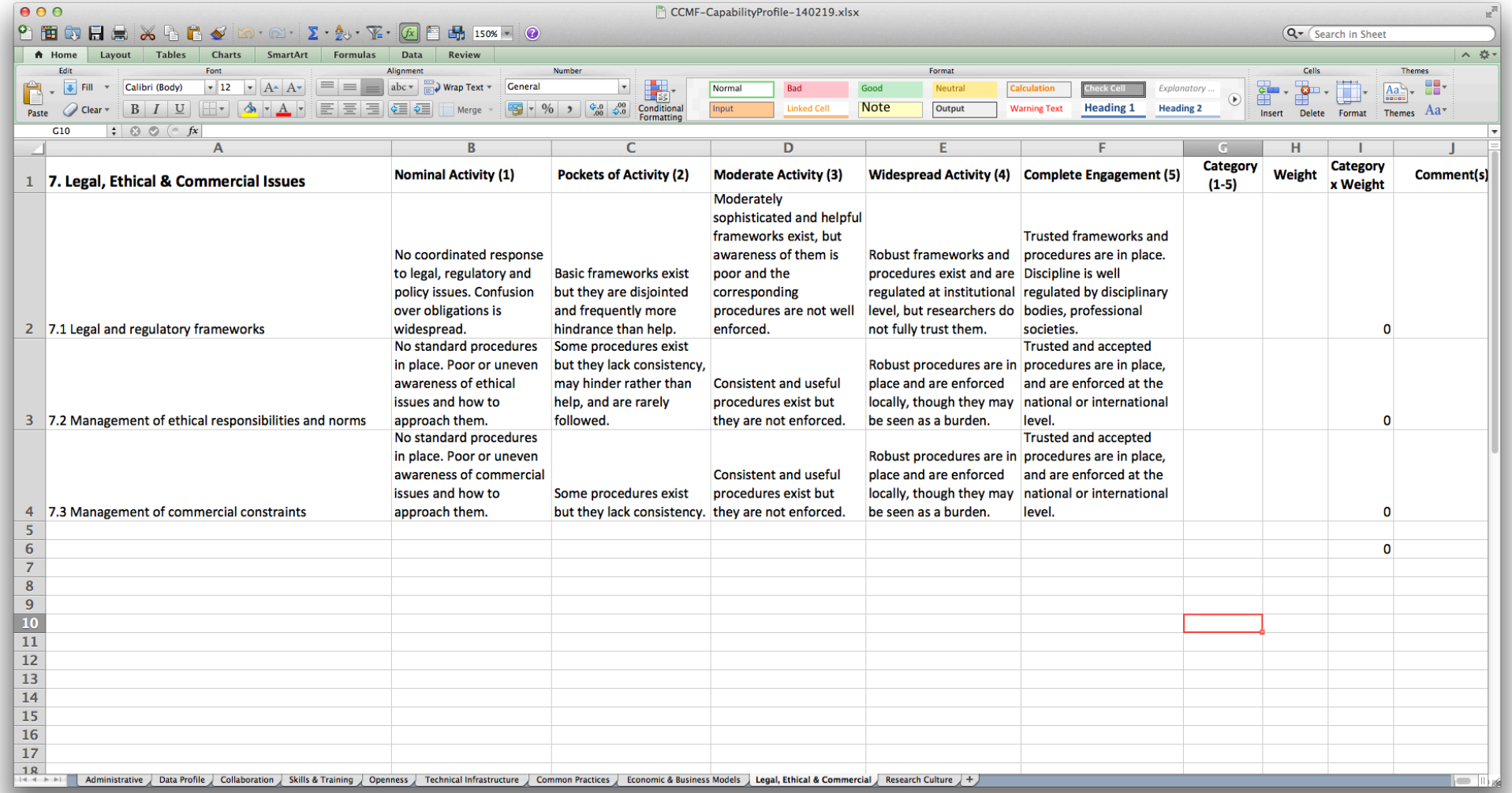
Administrative | Data Profile | Collaboration | Skills & Training | Openness | Technical Infrastructure | Common Practices | Economic & Business Models | Legal, Ethical & Commercial | Research Culture

Economic & Business Models



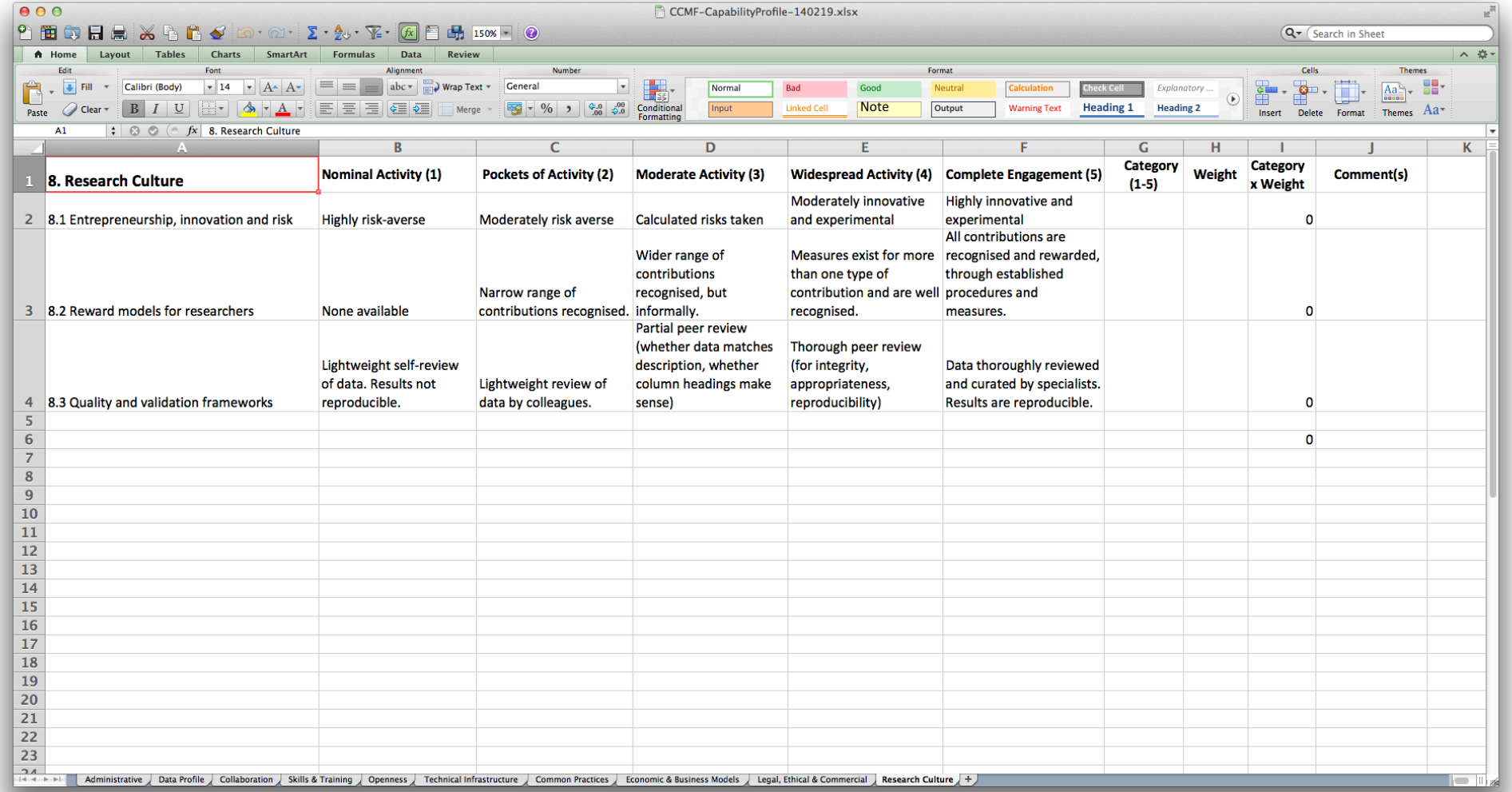
6. Economic & Business models	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)	Category (1-5)	Weight	Category x Weight	Comment(s)
6.1 Sustainability of funding for research	One-off funding focused on quick returns e.g. 1-2 years	Funding focused on short-term projects and quick returns e.g.2-3 years	Longer term investments on a 3-5 year timescale.	Single-phase thematic investments on a 5-7 year timescale.	Multi-phase thematic investments in 5-10 year blocks which build a community e.g. NSF DataONE Programme			0	
6.2 Geographic scale of funding for research	Projects funded internally.	Projects funded through grants from regional agencies.	Projects funded by national funders.	Projects funded by multiple national funders e.g. UK BioBank	Funding by international bodies and bi-lateral initiatives between national funders.			0	
6.3 Size of funding for research	Short investigative projects to encourage open innovation	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).	Large multi-national projects e.g. Virtual Observatory			0	
6.4 Sustainability of funding for infrastructure	One-off investments with no commitment to sustainment.	Multi-phase projects to develop infrastructure e.g. networks and services	Sustained multi-decade investments in data centres and services.	Infrastructure projects allowed slow transition to self-financing model.	Self financing infrastructure, networks and services			0	
6.5 Geographic scale of funding for infrastructure	Projects funded internally.	Investments by a single funding body at regional level.	Investments by a single funding body at national level.	Collaborative development at the national level by multiple funders e.g. Australian eResearch Orgnaisation (AeRO).	Collaborative development between international funders e.g. Elixir			0	
6.6 Size of funding for infrastructure	Small-scale tool development e.g. hackathons	Medium scale investments in networks and services e.g. Institutional Repositories	Co-ordinated investments in distributed systems.	Large central investments in network infrastructure or tools e.g. UK's JANET network	Large multi-national investments e.g. Large Hadron Collider			0	
		Informal collaboration with industry but no	Corporate / SME are non-funded partners in	Research is co-funded by industry and other	Established formal co-investment partnerships running long-term multi-				

Legal, Ethical & Commercial



	A	B	C	D	E	F	G	H	I	J
1	**7. Legal, Ethical & Commercial Issues**	**Nominal Activity (1)**	**Pockets of Activity (2)**	**Moderate Activity (3)**	**Widespread Activity (4)**	**Complete Engagement (5)**	**Category (1-5)**	**Weight**	**Category x Weight**	**Comment(s)**
2	7.1 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.			0	
3	7.2 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.			0	
4	7.3 Management of commercial constraints	No standard procedures in place. Poor or uneven awareness of commercial issues and how to approach them.	Some procedures exist but they lack consistency.	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.			0	
5										
6									0	
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										

Research Culture



8. Research Culture	Nominal Activity (1)	Pockets of Activity (2)	Moderate Activity (3)	Widespread Activity (4)	Complete Engagement (5)	Category (1-5)	Weight	Category x Weight	Comment(s)
8.1 Entrepreneurship, innovation and risk	Highly risk-averse	Moderately risk averse	Calculated risks taken	Moderately innovative and experimental	Highly innovative and experimental			0	
8.2 Reward models for researchers	None available	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.			0	
8.3 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.			0	
								0	