A Provenance Maturity Model

Kerry Taylor, Robert Woodcock, Susan Cuddy, Peter Thew, David Lemon

A Provenance Maturity Model

Kerry Taylor[1,2], Robert Woodcock[3], Susan Cuddy[4], Peter Thew[1] and David Lemon[4]

[1] *CSIRO Digital Productivity, Canberra, Australia*
[2] *Australian National University, Canberra, Australia*
[3] *CSIRO Mineral Resources, Canberra, Australia*
[4] *CSIRO Land and Water, Canberra, Australia*
firstname.lastname@csiro.au

Abstract. The history of a piece of information is known as "provenance". From extensive interactions with hydro-and geo-scientists in Australian science agencies we found both widespread demand for provenance and widespread confusion about how to manage it and how to develop requirements for managing it.
We take inspiration from the well-known software development Capability Maturity Model to design a Maturity Model for provenance management that we call the PMM. The PMM can be used to assess the state of existing practices within an organisation or project, to benchmark practices and existing tools, to develop requirements for new provenance projects, and to track improvements in provenance management across an organisational unit.
We present the PMM and evaluate it through application in a workshop of scientists across three data-intensive science projects. We find that scientists recognise the value of a structured approach to requirements elicitation that ensures that aspects are not overlooked.

Keywords: provenance · reproducibility · lineage · pedigree · requirements

1 Introduction

As the trend towards data-intensive cyberscience picks up pace, scientists are becoming increasingly concerned about data provenance. As consumers of data, scientists need to know: Where did this data come from? Is it good enough for me to use? Can I trust it? As producers of data and expert opinion, scientists need to ensure their results are scientifically credible, repeatable, and justified by the methods used and the reasoned interpretations made.

Knowledge of the history of a piece of information (where it came from, what it was generated for, and the workflow that generated it) is known as "provenance", although the terms "audit trail", "lineage" and "pedigree" are common synonyms. Understanding the provenance of a piece of information can be as important as the information itself. Using provenance, it should be possible to understand whether or not a piece of information is fit for the intended purpose or whether the information should be trusted.

While these general concerns are quite widespread, implementing software systems for long-term provenance management needs much more thought about requirements for the business context, capture, representation and storage, retrieval, and usability. From extensive interactions with hydro-and geo-scientists in Australian science agencies we found widespread confusion around how to move from high level descriptions of outcomes towards statements of requirements for selection of tools and methods for implementation. This problem is magnified in multidisciplinary science projects.

In this paper a novel tool for developing requirements for provenance is presented, the Provenance Maturity Model (PMM). The PMM has been developed together with a method for provenance requirements elicitation that is not further discussed here. It has been specifically developed for a context of multiple, disconnected stakeholders who are generally unaware of the drivers, challenges and tradeoffs of provenance management, but it can be useful in any scenario where an implementation of provenance management is required, such as hydrological modelling [15], agricultural research [3], emergency management [13] and chemistry lab notebooks [2]. The PMM can also be used to classify existing or aspirational tools and approaches to provenance to aid in tool selection after requirements are established.

Our development of the PMM was driven by our experience in the Bioregional Assessment program of the Australian Government; a program to understand the potential impacts of coal seam gas and large coal mining on water resources and water-related assets. It is a complex intergovernmental resource development decision-making process, with expectations of significant long term commercial and public interest in the decisions to be made. While there was very strong awareness of the necessity for provenance amongst the stakeholders, we struggled to discuss the breadth and depth of the impact of simple requirements statements. Elements of *maturity* have a major impact on the cost, distance from current practice, the ability to capture, and the ability to interoperate over organisational and disciplinary boundaries.

The content of the PMM was assembled by the authors' analysis of remarks and expectations of stakeholders in that and previous resource-exploitation scenarios. A survey of the research literature was also used to insert capabilities and maturity points that we may have missed. The PMM then provided a way to hold the conversations that were necessary to understand these issues where previously we had role, process and scope confusion. We needed the additional dimension of maturity to facilitate the conversation along with a context of cost-benefit and risk.

2 The PMM

Like the Capability Maturity Model (CMM) [12], the PMM contains a matrix of capabilities described at five levels of maturity. The capabilities are grouped into *Provenance Business*, *Data Management*, *Provenance Capture*, *Provenance Representation and Storage*, *Provenance Retrieval*, and *Usability*.

Table 1: PMM Capabilities(Rows)

PROVENANCE BUSINESS	This category relates to the adoption and commitment to provenance management. It also contains a grab bag of issues that need to be addressed for any particular provenance venture
Longevity	How long are provenance records designed to be kept?
Software maturity	What is the maturity of the software being used; for example does it have a known history of usage elsewhere or is it unknown to the organisation?
Organizational awareness: culture and behaviour	Is there a culture that understands and appreciates provenance?
Value recognition	Does the organisation care about provenance?
Governance	How formalised is the approach to provenance?
Perspective	The social extent of provenance sharing and interpretability

Transaction costs: Production vs. retrieval cost	How to balance the cost of detailed record keeping versus the cost of collection recording to provide an answer to a query
DATA MANAGEMENT	This category of capabilities relates not so much to provenance as to the management of the underlying data over which provenance operates. It is included here because certain features are necessary preconditions to provenance maturity
Data safety	The resilience of data (that is the subject of provenance) to failures in hardware and/or processing errors
Data versioning	Is there a scheme to identify data versions?
Data lineage	Is information kept to identify the source of data?
Identifier Management	Is the underlying data unambiguously identified?
Digital Preservation	How well preserved are the digital artefacts that are the subject of provenance?
PROVENANCE CAPTURE	This category refers to the processes, methods and tools for initial capture, or record-keeping, of provenance
Decision making and consultation	What records are kept on decisions made by people and groups and how integrated are these records to a provenance database?
Temporal scope (Start and end points)	In the scope of a process, when do we start capturing provenance and when does it end?
Granularity of capture	The smallest unit of a process that generates a provenance record - from a complete workflow down to individual executables and commands
Temporality/Currency	When was the provenance record made with respect to the execution of the process: during the process step-by-step, at specific stages or after the fact?
Software tools used in process	What information is captured on software used within a workflow & the workflow system itself; for example version number?
Hardware/platform	Information on the hardware environment when a workflow is executed
Provenance capture integration	How integrated is the collection of provenance records in the process execution? For example does the carrying out of the process generate provenance transparently or are the provenance records obtained by post-processing log files or independent data entry?
Provenance quality	How trustworthy is the provenance record that is kept?
Sophistication of automation	How automated is the capture of provenance record; for example manually or embedded into workflow processes
PROVENANCE REPRESENTATION & STORAGE	These capabilities discuss the static aspect of provenance information; in between creating it and using it
Provenance format	Do the provenance records adhere to a standard, and is that standard an international one or bespoke to the organisation?

Provenance language	How formalised is the content of the provenance records; for example free text or a strictly controlled language (with specific meanings)
Provenance security	Can the provenance records be validated or can they be corrupted or altered after collection; is there a way to determine if corruption has occurred?
PROVENANCE RETRIEVAL	These capabilities address the methods and support for obtaining provenance information
Provenance availability	How easy it is for people (or systems) to access the provenance records?
Provenance discovery	How easy is it to find the particular provenance record they need?
Licence to use provenance	Are the conditions of use of the provenance itself well understood?
PROVENANCE USABILITY	These capabilities refer to how provenance can be, or is, used: what is it all for?
Human readable	What attention is given to supporting the human interpretation of the provenance record?
Repeatable by automation	Can a workflow be repeated using the provenance records to identify all components and parameters that were used in the original workflow?
Reusable, that is, repeatable with improvement	Can a workflow be repeated using the provenance records but also allow deliberate substitutions; for example, an updated model or new dataset?
Transparent	Can you see what judgement decisions were made, and why?
Answerable (variation in results can be explained)	Can the provenance records identify the component that is the source of error or difference within a workflow with respect to an alternative?
Cross-disciplinary application	Is everyone talking the same language? Are they being forced inappropriately to talk the same language?

The five levels of maturity (columns) are labelled and described as follows where the original CMM title *Repeatable* has been replaced with *Tactical* to avoid conflicts with the use of that term in the study of provenance. The descriptions of each level have been modified to be more appropriate for provenance.

Initial(Chaotic) It is characteristic of provenance treatment at this level that it is (typically) undocumented and in a state of dynamic change, tending to be driven in an ad hoc, uncontrolled and reactive manner by users or events. This provides a chaotic or unstable approach to provenance management and provenance services and certainly implies that any services developed will be of very low functionality and scope.

Tactical It is characteristic of provenance treatment at this level that many aspects of data production and management are carried out with record-keeping in mind. Some attention is being made to ensure that identified processes are repeatable in some circumstances, possibly with consistent results. Discipline is unlikely to be rigorous, but where it exists it may help to ensure that following an audit trail is possible under stress.

Defined It is characteristic of provenance treatment at this level that there are sets of defined and documented standard processes established and routinely followed. These standard pro-

cesses both require and enable standard tools to be developed and used. Such tools both assist in the implementation of the processes and offer provenance services to derive value from provenance across the organization or community.

Managed It is characteristic of provenance treatment at this level that it has become uncontroversial: moved into the background as well-managed practices that are embedded in the fabric of business, travelling smoothly from project to project. It is consistently and effectively controlled and widely used.

Optimising It is a characteristic of provenance management at this level that the focus is on continually improving performance of provenance management itself as a lever for continuous improvement of performance of the underlying scientific or administrative processes, through both incremental and innovative technological improvements. Provenance is a highly valued component of business delivering transparency, accountability and knowledge management.

The full PMM matrix of 33 capabilities by 5 maturity levels is included at the end of this paper.

3 Evaluation

The PMM was used for requirements elicitation for three projects of two government science agencies in August 2013, over a two-day workshop. Background material was provided, including the PMM evaluation of some existing tools, many of which were known to the participants, such as ISO-19115-LE [6] and Prov-O [7]. The workshop included a half-day of presentations on the nature of provenance and some known tools and methods; brief presentations on the selected projects' goals; an introduction on the PMM and how to use it; a hands-on application of PMM to the three projects in project groups; a subsequent requirements documentation exercise; a plenary analysis of consequent requirements for the agency as a whole; and the joint development of an architecture sketch for the agency-wide provenance management. Those who were present for the full two days excluding the PMM developers and facilitators, that is 13 people, were surveyed at the beginning and again at the end of the workshop.

The survey participants were invited to respond to 21 questions of which 18 were phrased on a 5-level balanced Likert scale and all 21 included requests for free-text comments. In some cases participants indicated responses either in between or spanning consecutive levels; in all cases these have been treated as if the lowest (i.e. least positive) level was selected.

The lowest-scoring question overall was the pre-workshop question *"Please rate your familiarity in dealing with issues or tools relating to provenance (1=not all aware, 5=extremely aware)"*, for which the minimal response was 1, the maximum 5, and the average 3.2. The highest-scoring question overall was the post-workshop *"The approach to provenance management for your project will benefit by the application of the PMM as you have used it in the workshop (1=strongly disagree, 5=strongly agree)"*. The minimum response was 3 and the maximum was 5, with average 4.8. Other high-scoring questions referred to the contribution of the PMM towards developing user requirements. The lowest-scoring post-workshop question (average 3.7) referred to the ease of determining evaluation criteria, an element of the PMM application methodology that is out of scope for this paper. We can conclude that the participants found the PMM worthwhile.

The participants were also invited to suggest needs for clarification or other improvement to the PMM version that was used in the workshop. The PMM presented here has had some wording and sequencing adjustments since the workshop to take account of those suggestions.

4 Related Work

We have evaluated some existing tools and methods with respect to the PMM in order to provide some benchmarks to assist in interpretation of the PMM, and on the other hand, to assist project designers to locate tools that might help them achieve desired maturity levels. In this context, Prov-O [7] features due to its potential contribution towards high-maturity provenance Representation and Storage, therefore also contributing to high maturity Retrieval and Usability. Because of both its underlying flexible graph representation, and the ontology inference coupled with domain-specific ontology extensions, it is possible to use it to traverse widely differing domain-oriented provenance records through semantic, executable mappings to Prov-O, such as described in [3]. This can be done even when the primary record-keeping may be entirely ontology-unaware, such as is enabled by the mapping from ISO19115 Lineage to Prov-O [14]. An alternative ISO19115-driven extension to Prov-O has been developed [9] that would be useful when an early commitment to Prov-O is made. We can envisage the particular utility of the Prov-O graph representation to support dynamic provenance assembly for federated information systems like [16], [1] and [17] that support retrieval of data products or query-answering over compositions of resources. In unpublished work, we are also exploiting the inference capability to support arbitrary provenance comparison, building on the graph matching of [8].

Recent work on provenance for an integrated ecosystem approach to management of large marine ecosystems [5] demonstrates a high level of maturity for Capture, whereby a Web application for the development of data products and charts, tables, and map visualisations also keeps track of steps taken then embeds the provenance in the final PDF report. The related Global Change Information system will demonstrate a high level of maturity for Data Management, particularly for identifier management over a heterogenous contributor community [10].

The notion of *Research Objects*[11], especially computational research objects [4], contributes to a very high level of maturity in Usability, with provenance records very closely tied to executable components for repeatability and reusability and also to the scientific practice.

5 Conclusion

We present the Provenance Maturity Model as a part of a structured approach to developing requirements for provenance management in data-intensive science. The PMM lays out many characteristics of provenance management in a matrix where preferred options may be considered and selected in the context of some evaluation criteria. We found that scientists recognise the value of a structured approach to requirements elicitation that ensures the depth and breadth of the issues are considered and that aspects are not overlooked. The value of the framework in clarifying the language in an "industry standard" approach is also appreciated.

We recommend using the PMM in a workshop environment once the scientific content of a project is well enough understood to commence. We recommend developing an evaluation criteria then proceeding to check cells in the PMM by consensus, adding additional rows if necessary. Later, system and software requirements can be developed in a conventional way, with frequent reference to the instantiated PMM. A record of tools and business processes that have previously been benchmarked by the PMM can help to fill in a solution architecture. Much later, the PMM can help to review provenance goals and to consider advancing the maturity.

In future work we would like to re-evaluate the PMM in an alternative cultural, organisational and problem context, and also to track the influence on project results through project life-cycle case studies.

Acknowledgement The authors thank the anonymous reviewers and the many collaborators for their questions and insight, especially Neal Evans and Brian Hanisch of Geoscience Australia.

References

1. R. Ackland, K. Taylor, L. Lefort, M. Cameron, and J. Rahman. Semantic service integration for water resource management. In Benjamins Gil, Motta and Musen, editors, *The Semantic Web–ISWC 2005: 4th International Semantic Web Conference, Galway, Ireland*, volume LNCS Volume 3729 / 2005, pages 816 – 828. Springer, November 6–10 2005.
2. Nico Adams, Armin Haller, Alexander Krumpholz, and Kerry Taylor. A semantic lab notebook—report on a use case modelling an experiment of a microwave-based quarantine method. In *Linked Science (LISC2013)*, volume 1116. CEUR proceedings, October 2013.
3. Michael Compton, David Corsar, and Kerry Taylor. Sensor data provenance: SSNO and PROV-O together at last. To appear 7th International Semantic Sensor Networks Workshop, October 2014.
4. David De Roure. Towards computational research objects. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts*, DPRMA '13, pages 16–19, New York, NY, USA, 2013. ACM.
5. M. Di Stefano, P. Fox, S. Beaulieu, A. Maffei, P. West, and J. Hare. Enabling the integrated assessment of large marine ecosystems: Informatics to the forefront of science-based decision support. In *AGU Fall Meeting*, number Poster IN51A-1689, San Fancsisco, December 2012. American GeoPhysical Union.
6. ISO 19115-2:2009 geographic information - metadata - part 2: Extensions for imagery and gridded data. ISO19115-2 Standard, 2009.
7. Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. PROV-O: The PROV ontology. W3C Recommendation, 2013. Available at `http://www.w3.org/TR/prov-o/` (accessed 23^{rd} April 2014).
8. Qing Liu, Xiang Zhao, Kerry Taylor, Xuemin Lin, Geoffrey Squire, Corne Kloppers, and Richard Miller. Towards semantic comparison of multi-granularity process traces. *Knowledge-Based Systems*, 52:91–106, November 2013.
9. Francisco J. Lopez and Jesus Barrera. Linked map VGI provenance schema. Deliverable D1.6.1, Planet Data Network of Excellence, March 2014.
10. X. Ma, P. Fox, C. Tilmes, K. Jacobs, and A. Waple. Capturing provenance of global change information. *Nature Climate Change*, 4(6):409–413, 2014.
11. Kevin Page, Raúl Palma, Piotr Hołubowicz, Graham Klyne, Stian Soiland-Reyes, Don Cruickshank, Rafael González Cabero, Esteban Garciá Cuesta, David De Roure, Jun Zhao, and José Manuel Gómez-Pérez. From workflows to research objects: An architecture for preserving the semantics of science. In *2nd International Workshop on Linked Science 2012: Tackling Big Data (LISC2012)*, volume 951, Boston, USA, November 2012. CEUR Proceedings.
12. M.C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber. Capability maturity model, version 1.1. *IEEE Software*, 10(4):18–27, 1993.
13. R. Power, C. Wise, B. Robinson, and G. Squire. Harmonising web feeds for emergency management. In J. Piantadosi, R.S. Anderssen, and J. Boland, editors, *MODSIM2013, 20th International Congress on Modelling and Simulation*, pages 2194–2200. Modelling and Simulation Society of Australia and New Zealand, December 2013.
14. Yanfeng Shu and Kerry Taylor. ISO 19115 lineage ontology. Online, January 2013. Accessed November 2013.
15. Yanfeng Shu, Kerry Taylor, Prasantha Hapuarachchi, and Chris Peters. Modelling provenance in hydrologic science: A case study on streamflow forecasting. *Journal of Hydroinformatics*, 2012.
16. Kerry Taylor, Tim Austin, and Mark Cameron. Charging for information services in service-oriented architectures. In *Proceedings, International Workshop on Business Services Networks (BSN 2005), Workshop of IEEE International Conference on e-Technology,e-Commerce and e-Service*, pages 112–119, Kong Kong, March 2005.
17. Robert Woodcock, Bruce Simons, Guillaume Duclaux, and Simon Cox. AuScope's use of standards to deliver earth resource data. In *Geophysical Research Abstracts*, volume 12:EGU2010-1556. European GeoPhysical Union General Assembly, 2010.

Table 2: The PMM capabilities × Maturity levels

Provenance Business *This category relates to the adoption and commitment to provenance management. It also contains a grab bag of issues that need to be addressed for any particular provenance venture.*

Capability	**Initial (Chaotic)**	**Tactical**	**Defined**	**Managed**	**Optimising**
Longevity	No provenance	Lifetime of a process instance execution	Lifetime of an identified problem or project to which knowledge is being applied	Lifetime of a consequent action plan or agreement	Deliberately unbounded without prescription in advance; i.e. evolutionary
Software maturity	Tools are not aware of provenance	Experimentally developed	Extensively trialled	Several alternative implementations available, at least some are robust	Appropriate products are available and are also supported and maintained
Organisational awareness: culture and behaviour	No support for provenance	Individual or small team initiative	Major organisational support for trial or development but lacking maintenance commitment	Policy; Long term commitment supported by explicit funding stream	Legislation or Regulation
Value recognition	No support for provenance	Value of provenance is recognised in the custodian (asset)	Value of provenance is exploited opportunistically	Provenance value exploitation is intrinsic part of business model	Provenance is recognised as a knowledge base of evolving scientific method and used for continuous improvement processes
Governance	No support for provenance	Implicit (everybody knows it, but it is not formally captured)	Explicit. Written, formal, contractual management	Provenance record management plans are comprehensive and followed	Formal accountability and governance throughout lifecycle established with & between parties
Perspective	No support for provenance	Individual: I know what I did	Team: We know what we did	Organisation: we know what our teams did	Community: we all know what everyone did

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Transaction costs: Production vs. retrieval cost	No support for provenance	Low cost of production but retrieval is very expensive and slow and not always possible	Trade-off recognised and accommodated in many cases	Incremental cost of retrieval close to zero where justified	Identified cost / benefit habitually used in design decisions

Data Management *This category of capabilities relates not so much to provenance as to the management of the underlying data over which provenance operates. It is included here because certain features are necessary preconditions to provenance maturity*

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Data safety	No backups	Data is manually backed up	Data backup is part of an automated backup process	Data backups are stored off-site	Data backups undergo regular restore tests
Data versioning	Unknown versions	Old data replaced by new data. Version identification associated with new data	Old data is archived and associated with version identification.	Version control of collections	Data elements (e.g database tuples) dated and annotated with provenance
Data licensing	Unknown	Licence conditions are recorded, or a system wide standard licence applies	Licence need not be system-wide and conditions are retrievable from point of access to data	Licence need not be system-wide and conditions are retrievable from point of access to data	Flexible licencing policies can be expressed and computationally validated in the context of the intended use of the data
Identifier Management	None	Locally-scoped identifiers are assigned, possibly through a file-naming convention	Systematic, unique identifier assignment, crossing datatypes and technology platforms, supporting retrieval of identified objects	Systematic, globally unique and resolvable identifier management	International standards followed; resolvable identifiers maintained over time

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Digital preservation	None	Some project materials and output data are archived	Systematic data preservation mechanisms are in place for identified strategic data	Digital preservation strategies extend to digital artefacts such as software and minuted decisions	Digital preservation strategies follow international standards and are regularly reviewed for scope, best practice, and longevity

Provenance Capture *This category refers to the processes, methods and tools for initial capture, or record-keeping, of provenance*

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Decision making and consultation	No records	Decision-taking meetings and individual choices in process are documented and justified	Documentation retrievable from multiple access points	Judgement choices are entirely transparent and fully integrated into provenance record and services	Reasoning services incorporated e.g. to identify decisions that follow policy (or not)
Temporal scope (Start and end points)	No provenance	Case-by-case; driven by other procedural concerns	Defined and applied at project level with end purposes in mind	System-wide principles for provenance established that determine scope	Generally lifecycle-complete, but entirely adaptable to cases without loss of verity
Granularity of capture	Undetermined	Coarse grained. Resources and methods are loosely described	High-level components captured but not all well described in terms of role or properties	Low-level components captured from the point of view of decisions made, steps taken, tools and data used	Granularity of capture is driven by understanding of future requirements
Temporality / Currency	Provenance not captured	Reconstructed on demand, after the fact	Constructed on demand from contemporary notes	Real time production but decaying record	Designed for long term storage and interpretation

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Software tools used in process	Untraceable	Identified by commonly understood monikers	Versions, dates and providers rigorously identified	Originals archived with descriptions, including metadata	Fully integrated into provenance services
Hardware / platform	Untraceable	Identified by commonly understood monikers	Versions, dates and providers rigorously identified	Originals archived with descriptions, including metadata	Fully integrated into provenance services
Provenance capture integration	No integration	Provenance is created by separate processes, usually running in parallel to workflows	Integration into tools and workflows	Standards based capture from tools and workflows	Plug and play with whatever is needed, tracking provenance
Provenance quality	Unknown	Unreliable or partial; user feedback collected and published	Measured occasionally; quality may be inferred from other attributes such as author, date	Measured routinely, quality limits and impact are understood	Effective methods to detect and improve bad provenance in place. All provenance is trustworthy; or trustworthiness is well documented
Sophistication of automation	None / manual	Policies for collection are established and followed but interpretations are localised	Habitual recording in a systematic (e.g. tablular) way in identifiable documents that are systematically archived and validated	Capture is embedded in data- and decision-processing software; some aspects demanding operator input	Integrated into the culture and toolsets

Provenance Representation & Storage *These capabilities discuss the static aspect of provenance information; in between creating it and using it*

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Provenance format	No provenance	Provenance may be mentioned in key papers; simple schemes like filename conventions and time stamps may be used	Formal standard for provenance format adopted and practised at key places	Flexible standard or tool-dependent formats prescribed and followed according to minimal capture granularity	Interoperable mappings over multiple formats implemented; may rely on overarching standard
Provenance language	None	Interpretation relies on natural language methods	Information is captured by link to controlled vocabulary with glossary	Semantics is captured by link to a formal ontology	Representations are interpreted for interoperability and adaptation to context of use
Provenance security	None	Original provider of provenance identifiable	Formal processes for authentication and audit trail	Provenance is signed and tamper-proof, within an organisation on selected transactions	Non-repudiation

Provenance Retrieval *These capabilities address the methods and support for obtaining provenance information.*

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Provenance availability	Clueless	Non-automated; requires judgement	Database or web page	Web Service	Direct availability to analysis and reporting tools; API supporting structured queries

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Provenance discovery	"Phone a friend"	Retrievable from point of data product identification	Full text search over provenance records-can retrieve corresponding data through provenance search	Search by provenance structure and components: can retrieve corresponding data	Search for provenance patterns: can analyse provenance itself as subject of enquiry
Licence to use provenance	Unknown	Licence conditions are recorded or a system-wide standard licence applies	Licence need not be system-wide and special conditions are retrievable from point of access to provenance	Privacy or confidentiality conditions on access to provenance are enforced by tools	Flexible licencing policies can be expressed and validated computationally at the point of access to provenance

Provenance Usability *These capabilities refer to how provenance can be, or is, used: what is it all for?*

Capability	Initial (Chaotic)	Tactical	Defined	Managed	Optimising
Human readable	No	Only	Predominantly, with some machine-processable structure	Detailed, but detail can obscure meaning	Presentation tools take account of user perspective and purpose
Repeatable by automation	No	There is some chance that automated sub-processes are repeatable with considerable investment of effort	Partially automated to the extent that the general method can be reapplied for (typically) different results	Fully automated, repeatable results is possible in some cases	Processes are automatically repeatable
Reusable, that is, repeatable with improvement	Opaque	There is some chance (decreasing over time) that sub-processes are repeatable with considerable investment of effort	Editing of parameters or selected data for rerun is supported	Processes may be arbitrarily edited or built upon or varied for improved results	Patterns in provenance are discernible and used for process improvement

Capability	**Initial (Chaotic)**	**Tactical**	**Defined**	**Managed**	**Optimising**
Transparent	Opaque	Only within small project teams	Formalised approach to transparency ensures that some decision points are noted and justified	Methods and tools are unambiguously identified but may not be interpretable by interested parties	Open access to justified methods and tools for all nominated parties; explanatory capability
Answerable (variation in results can be explained)	Impossible to say	Some sub-processes may be examined to explain variation but confidence is low	Variation can be attributed to plausible differences based on managed time-stamps or versions	Failure to reproduce can be diagnosed to identifiably different components	Automated; diagnosis limited only by original provenance collection granularity; user feedback quality included
Cross-disciplinary application	Unsuitable	Relies on serendipity	Provenance is available in a widely-used format that may be partly accessible to multiple discipline areas; generally relies on a lowest-common-denominator approach	Provenance is made available through multiple portals or in multiple formats to suit different discipline areas	Provenance management works for complex and intractable cross-disciplinary problems; Inter- or trans-disciplinary working is deliberately supported through management of multiple viewpoints