IJDC | *General Article*

# Using Metadata Actively

Colin L. Bird
University of Southampton

Simon J. Coles
University of Southampton

Iris Garrelfs
University of the Arts, London

Tom Griffin
STFC

Magnus Hagdorn
University of Edinburgh

Graham Klyne
Nine by Nine

Mike Mineter
University of Edinburgh

Cerys Willoughby
University of Southampton

**Abstract**

Almost all researchers collect and preserve metadata, although doing so is often seen as a burden. However, when that metadata can be, and is, used actively during an investigation or creative process, the benefits become apparent instantly. Active use can arise in various ways, several of which are being investigated by the Collaboration for Research Enhancement by Active use of Metadata (CREAM) project, which was funded by Jisc as part of their Research Data Spring initiative. The CREAM project is exploring the concept through understanding the active use of metadata by the partners in the collaboration. This paper explains what it means to use metadata actively and describes how the CREAM project characterises active use by developing use cases that involve documenting the key decision points during a process. Well-documented processes are accordingly more transparent, reproducible, and reusable.

*Accepted* 24 February 2016

Correspondence should be addressed to Dr Simon J. Coles, University of Southampton, Highfield, Southampton SO17 1BJ. Email: S.J.Coles@soton.ac.uk

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Copyright rests with the authors. This work is released under a Creative Commons Attribution (UK) Licence, version 2.0. For details please see http://creativecommons.org/licenses/by/2.0/uk/

 doi:10.2218/ijdc.v11i1.412

# Introduction

We all *use* metadata, of course we do, but how often do we do more than collect and save metadata? How often do we use metadata actively to guide future actions?

Gathering metadata can be seen as a burden imposed for the sake of archiving data, whereas enabling active use of metadata gives researchers advantage *during* their work. Active use can arise in various ways, such as informing decision steps in a process, feedback in a workflow situation, or iteration to obtain a better outcome. The following case provides a very basic illustration of using metadata actively to calibrate future activities.

Consider the process of cooking food in an oven, which we have set to a particular temperature. However, we are suspicious about the oven setting, because previous dishes have been overcooked, so we insert an independent temperature probe, which shows us that the actual oven temperature is 20 degrees higher than the setting indicates. We now have metadata that we can use actively to offset the oven setting and obtain the oven temperature that we require. While this example might seem trivial, we can generalise it for any process that produces outcomes subject to certain conditions: when we alter those conditions according to the metadata, we are using the metadata actively.

All processes generate metadata, some of which is captured, some not. A proportion of that metadata is purely descriptive, for example, date, time, and researcher identity. Such information can usually be recorded with the Dublin Core Element Set[1] and will be associated with any data that was generated by the process and is being preserved in a repository. Descriptive information of this nature is sometimes wryly described as "tombstone metadata". Other metadata is dependent on the context, so might record the conditions of an experiment or characterise the data that the process generated, which enables the discovery of that data for reuse.

Some metadata comprises values that can inform future steps and processes; act iteratively or as feedback to refine an output. For example, some processes involve an evaluation of the result(s) and some reasoning on consequent alterations of the metadata values in the subsequent process or processes. Use or reuse of such values constitutes active use of the metadata, which is the subject of this paper.

Although we are aware of the blurred and shifting border between metadata and data, we do not seek to argue the point. It is reasonable to question whether an element is data or metadata if it influences a process and, as an outcome of that process, is modified by the researcher or by an associated system, such that a succeeding process or a rerun of the same process is influenced differently. Most, but perhaps not all, scientists and engineers would opt for that element being 'metadata'. The situation can be much less clear for artists, some of whom experience a sense of vagueness surrounding the nature of the metadata that they would want to capture as artists, bearing in mind the *active use of metadata*, including reuse, and the divergence of practices, both within art and between art and science.

Our intentions are firstly to recognise situations where metadata could and should be used actively, then secondly to characterise the active use. This characterisation is likely to vary across and even within domains. To enable other researchers to benefit from exemplars of active use, and thereby enhance their research capability, some recognition of differences as well as commonalities would be necessary.

1 Dublin Core Metadata Element Set, Version 1.1: http://dublincore.org/documents/dces/

In this paper, we report our investigations under the auspices of the CREAM project, which is funded by Jisc[2] as part of their Research Data Spring initiative[3]. CREAM stands for *Collaboration for Research Enhancement by Active use of Metadata* and involves the University of Southampton, the University of Edinburgh, the University of the Arts London, the Science and Technology Facilities Council (STFC), and Nine by Nine. The partners have merged their experience in the support of research processes to collaborate in exploring how metadata can be used actively for the purpose of enhancing research.

Nine by Nine offers the software platform, Annalist[4], which has been our primary vehicle for modelling metadata. Both the University of Southampton and Nine by Nine have extensive experience in dealing with data and metadata from a very broad range of disciplines through generic and specific software developments. Southampton has specific expertise in digital research platforms, such as LabTrove (Milsted, Hale, Frey, and Neylon, 2013). STFC brings in several large centralised experimental facilities (8000 users) and as such has a highly evolved data management system for well-understood experimental processes conducted using large-scale analytic facilities (Matthews, Sufi, Flannery, Lerusse, Griffin, Gleaves, and Kleese, 2010). The University of Edinburgh's GeosMeta project focuses upon allowing researchers to flexibly define data about research activities and entities, for holding in a document-oriented database. In particular, users of software scripts can annotate these to record better the process by which their files are created. As well as covering a wide range of scientific practices, we also engage with projects led by the University of the Arts London (UAL) to ensure that the proposed principles of the active use of metadata are also applicable to arts and humanities subjects. This aspect of our collaboration enables an understanding of the extent to which art practice research varies from scientific research, for example, less rigid prior constraints and consequently a greater focus on using a retrospective provenance recording as a starting point. We have also observed similarities in the practices of science and the arts, notably in how decisions are made: many appear tacit or implicit, and are not recorded explicitly. Such observations have led us to emphasise the importance of documenting decisions and the information that informs them: a well-documented process is more transparent.

A key facet of the CREAM vision is to ensure that *metadata means more*. To enable research communities to realise this vision, we are working towards prototype tooling and guidance that will enable researchers, during a project or task, to collect and *use* metadata that informs future decisions.

## Active Use of Metadata

We have adopted the following provisional definition-cum-description of the active use of metadata:

> 'Metadata that is used actively comprises the specific assemblage of metadata and annotations that informs decisions made within a project or process and is capable of being reused within that project or by another process.'

2 Jisc: https://www.jisc.ac.uk/
3 Research Data Spring: https://www.jisc.ac.uk/rd/projects/research-data-spring
4 Annalist linked data notebook: http://dx.doi.org/10.5281/zenodo.44381

 doi:10.2218/ijdc.v11i1.412

One of the objectives of our work is to use this definition as the basis for providing guidance supported by exemplars, which will enable researchers in a wide range of disciplines to recognise, characterise, and exploit the metadata that they use actively. Figures 1 and 2 illustrate the concept of active use in the context of a generic procedure.

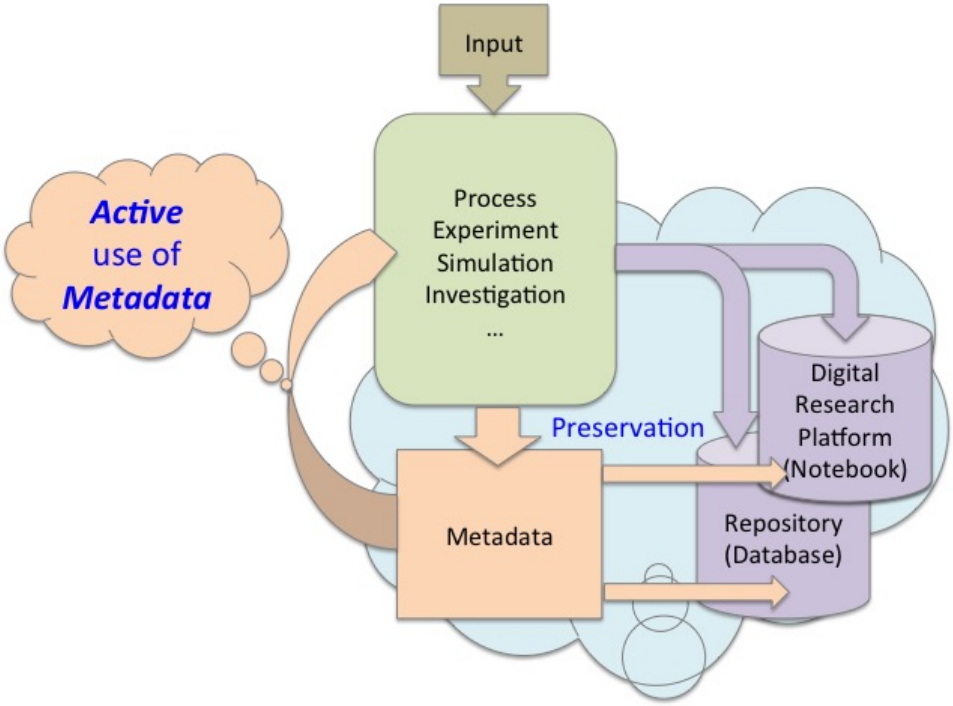


**Figure 1.** This diagram illustrates the repetition of a process using the same or modified metadata and also demonstrates preservation of the metadata in a repository (such as a database) or with a digital research platform (such as a notebook), thereby enabling that metadata to be reused subsequently or to be extracted by another process, which might be run by a different researcher.

Figure 3 is an example of the use of GeosMeta being tested in climate modelling and applications of micro-Xray tomography. GeosMeta permits researchers to capture at source what they have done when running software, by holding metadata such as parameter values, input and output filenames, information on the executing computer, software versions, etc. Active use of the metadata includes flagging that a file should not be reused because it is in error; discovering the processing chain that led to the existence of a particular file; seeing where a script or a data file has been used; finding where a particular parameter value has been used; capturing the status when an error is discovered; and inferring consequences for the validity of downstream data.

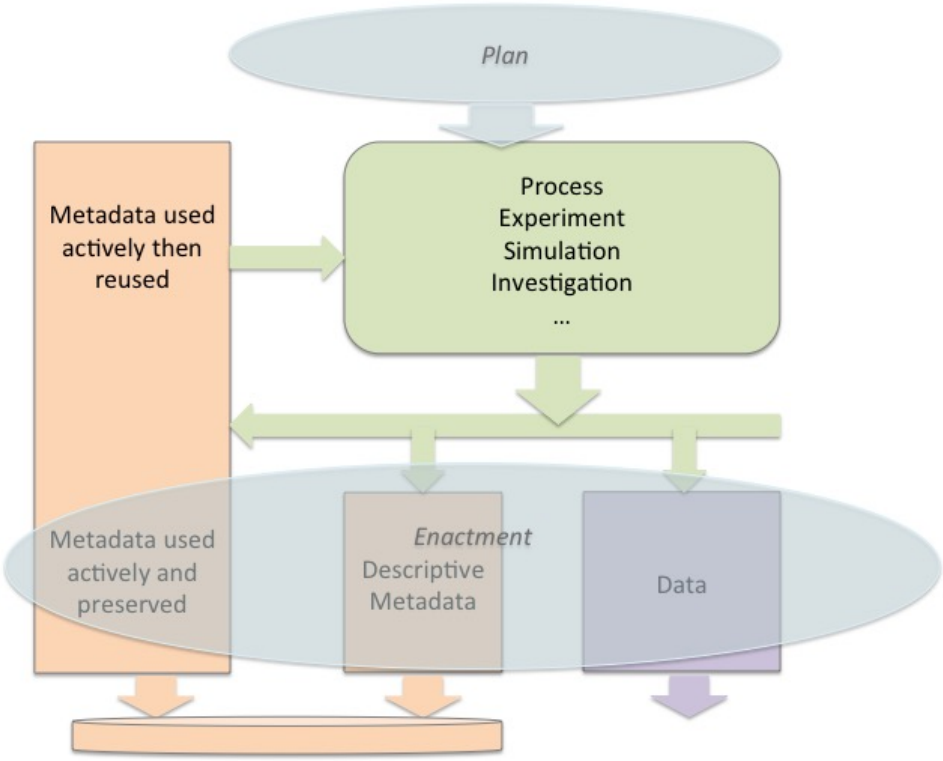


**Figure 2.** This diagram illustrates how metadata used actively has a role in the transition from planning to enactment, as can be represented by the prospective and retrospective provenance respectively.

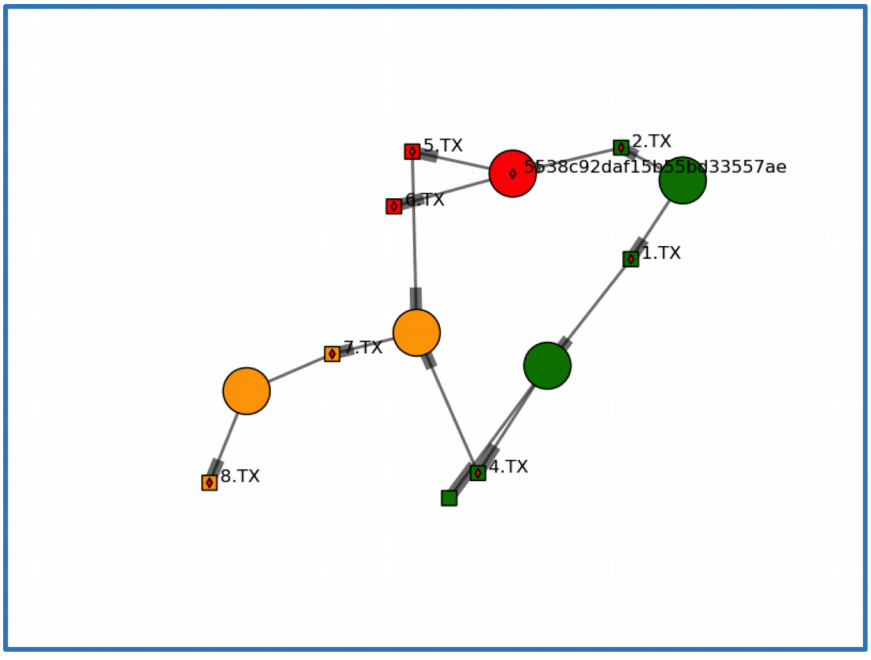


**Figure 3.** An example of the use of GeosMeta to permit researchers to capture what they have done when running software. The circles represent a document in the database. Each document holds metadata gathered during the execution of a script that takes input files and generates output files. The files are represented by the squares. The metadata are gathered by instrumenting the script with additional lines of code.

 doi:10.2218/ijdc.v11i1.412

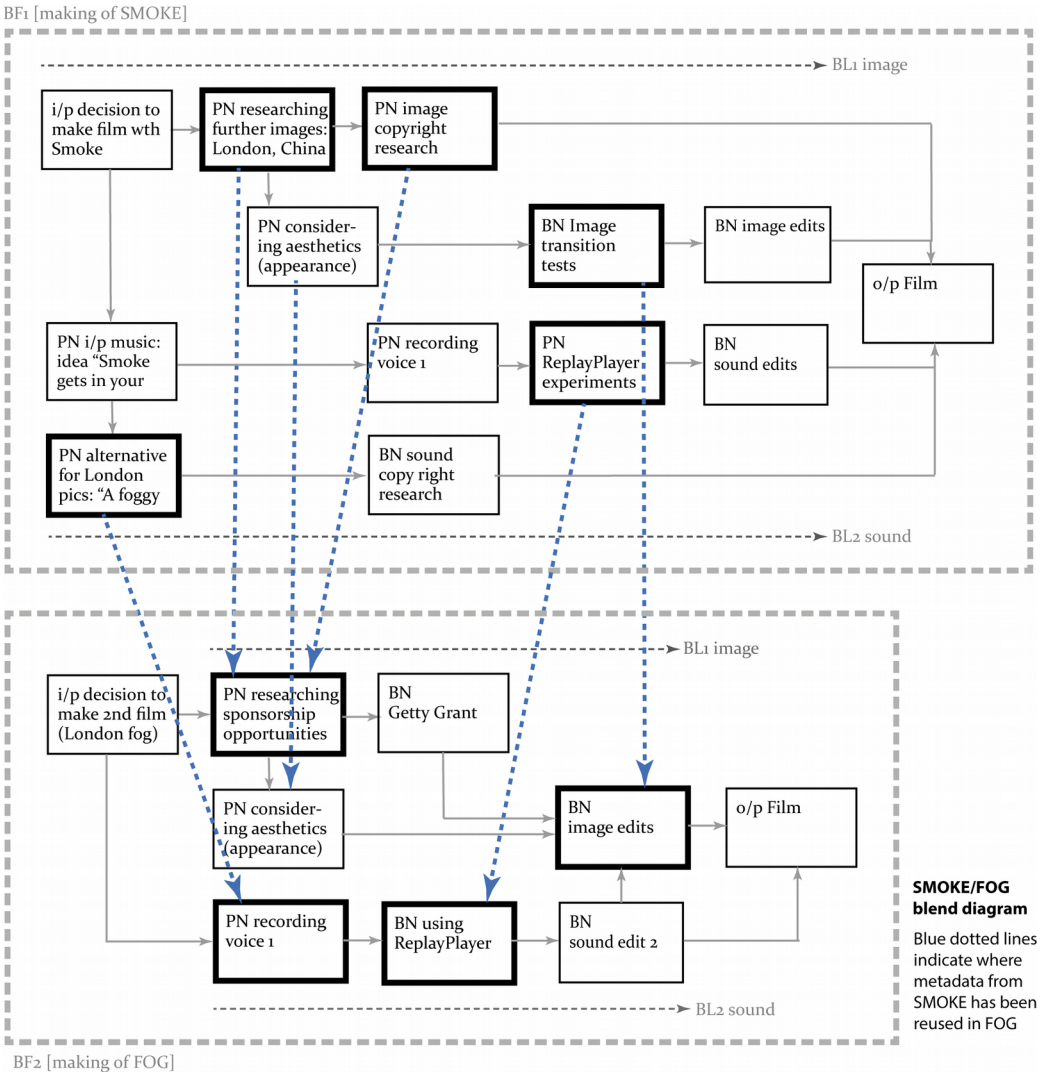


**Figure 4.** A blend diagram using concepts from Procedural Blending (Garrelfs, 2015). The upper part illustrates the creation of the audio-visual piece "Smoke"; the lower part relates to a second audio-visual piece, "Fog", for which the work began after the completion of "Smoke". The boxes depict Blend Nodes, with those highlighted in Smoke having influenced the later work; while those highlighted in Fog were influenced by the earlier work. The dotted blue lines show where the creator used Smoke metadata actively when making blending decisions while planning the Fog video.

Figure 4 is an application of Procedural Blending, a model of process in sound art practice developed by Garrelfs (2016) that aims to expand the discipline's discourse by considering process through an exploration of artists' perspectives, based on concepts from Conceptual Blending, a theory of cognition developed by Fauconnier and Turner (2002). Metadata generated in the upper model is used actively to influence the development of the lower model.

The GeosMeta and Procedural Blending scenarios demonstrate that active use transcends not only discipline but also the nature of the process involved. The CREAM project is also proposing to develop other exemplar scenarios: for experiments conducted at the large-scale STFC facility; for synthetic chemistry, interfacing with a digital research platform; and for structure determination using X-Ray Crystallography.

## Characterising Active Use

During the first phase of the CREAM project, our methodology was to model the example applications provided by the partners, aiming to extract the core elements of metadata that were being, or were capable of being, used actively. We intended to explore existing schema or vocabularies with which to represent the core elements, and to investigate vocabulary and packaging formats that we could use for metadata exchange. Initially, we expected work in the area of provenance capture and description to offer a good fit for describing process and the active use of metadata, even though it might be perceived differently in different domains. Indeed, looking for provenance information was a feature of our preliminary analysis of representative data sets, as described subsequently in this section. However, it transpired that we would need to go beyond a basic provenance framework; even vocabularies that represent the planning and enactment depicted in Figure 2 would not wholly encompass active use.

It will be apparent from the preceding sections of this paper that our methodology needed to evolve with experience. To begin with, we had expected to find common practice among the project partners, but in actuality we found very little, as the majority of the metadata in use was domain-specific. Although we did find that Dublin Core-like annotations were present in all the examples, such values represented neither the nature nor the means by which the metadata was being used actively.

Acknowledging the similarities and differences between the different domains, we adapted our methodology to focus more on recognising the factors that characterise the active use of metadata, i.e., how it is used. We expect also to take into consideration why metadata is being used actively, as we believe that to be an aspect of provenance.

Decisions are often based on tacit knowledge, and it can be difficult for others to understand the decisions that have been taken, and hence to follow the possibly novel path that has been discovered. At other points, knowing about decisions can aid us to explore alternative routes.

Sometimes the tacit knowledge may be so ingrained that we might not even realize that we are actually making decisions. However, when we become aware of and articulate those key decisions, it is easier to revisit them and to be more agile in our research. Furthermore, by recording the active use of metadata, we reveal more information about our process, and hence help others to interpret, validate, reproduce and reuse our work, generally increasing its overall value to science and art.

We believe that knowledge mining will be transformed by acknowledging the concept of the active use of metadata – so much around provenance and invention is inferred or assumed about a piece of information, often without any basis. Transforming information into knowledge implicitly requires an understanding of context and the rationale behind its creation.

Metadata modelling is a concept comparatively familiar to software engineers and system designers, but is less well understood in the sciences and to an even lesser extent in the arts. Models differ within as well as between disciplines, and the schema that represent the models are frequently difficult to map from one to another. Some processes are formulaic and thus may be ‘scriptable’, but most are rather more organic. Moreover, active use is inherently dynamic: the metadata elements do not necessarily conform to any established schema or pattern. There is a dearth of tools capable of characterising processes that involve active use of metadata, and a complete lack of tools capable of comparing processes.

 doi:10.2218/ijdc.v11i1.412

We used Annalist as our primary tool for analysing representative sets of data records, together with their associated metadata. We did this by taking descriptions and data from each of the contributing partners and constructing linked data models, using existing linked data[5] vocabularies (such as PROV[6]) where there was an obvious fit. The original expectation was that we would be able to identify and use some common terms across all or many of the examples. The main conclusion from this activity is that we have a good understanding of the metadata used in the processes that we have studied, but have experienced wide variations in the forms of capture of that metadata. For example, the data management system at STFC captures metadata automatically, whereas observations of chemistry research show that much of the true actively used metadata is captured in a narrative style of recording rather than in a structured format, although both are useful and important. In the chemistry context, we have begun to investigate how to extract the metadata by employing text mining and semantic analysis using workflows.

The creation of models with Annalist has also highlighted the need for complementary tools to enable active use in activities such as discovery, exploration, visualisation, reuse, remixing, integration, and curation for both the original researcher and for future collaboration and publication. We view identifying the use cases and exploring these functions as extremely important as a future aim for this project.

We are currently using the Annalist tool[7] to capture and characterise the data and metadata generated by these activities, aiming to explore a model in which a research process is defined, metadata is collected within that process, and subsequent actions in the research process defined within the template are determined by the metadata collected so far, thus determining a self-adaptive workflow that can be human-mediated. However, we remain very conscious that in creative practice direct responses to the unfolding process occur to varying degrees, such that in some instances reflexive documentation might be vital if metadata is to be captured. Moreover, practitioners make choices using variable criteria, not all of them conscious or reasoned, such as like or dislike. For scientists such changes can also occur during the process in response to observations or unexpected events.

For both scientists and artists, capturing the choices that are made and the decisions surrounding them is likely to be valuable for future understanding and improving processes in the future.

## Research Enhancement

A fundamental aim of the CREAM project is to enhance research by improving the recording of research processes in all disciplines, be they scientific or arts-based. De Roure (2010) suggested twelve Rs of the e-Research Record, of which the following seven properties of a research description are particularly relevant to our goal of enhancing research: repeatable, reproducible, reusable, repurposable, reliable, retrievable, and refreshable. Better documentation of procedures is essential for achieving the aims represented by these properties, with repeatability, reproducibility, and reusability being the foremost.

5 Linked data: https://www.w3.org/standards/semanticweb/data
6 Overview of PROV: https://www.w3.org/TR/prov-overview/
7 Annalist software and documentation: https://github.com/gklyne/annalist

Rigorous recording of procedures hinges particularly on documenting key decision points and all the information that guides the overall process, primarily the information considered to be actively used metadata. Well-documented procedures have the following benefits:

- They allow other researchers to follow the decision-making process, thereby speeding up their evaluation of data;
- They allow other researchers to insert new decision points along a workflow;
- They enable selective reuse of parts of a dataset;
- They increase transparency.

To date, we have developed a broad understanding of the characteristics and scope of the active use of metadata. In particular, one form that is commonly overlooked in conventional method documentation is the representation of the tacit knowledge that researchers have used to guide their decisions. The capture of tacit influences is an aspect to which the CREAM project has paid particular attention. Among the options under consideration is the practice of reflective documentation, thus helping achieve more effective documentation of procedures entailing the use of tacit knowledge.

Our mission going forward is as follows:

> 'Based on a portfolio of use cases that characterise the active use of metadata, supported by a toolset comprising data management and visualisation tools, to enhance the recording of research procedures, notably decision points, thereby to make processes substantially more repeatable, reproducible, and reusable.'

This mission statement will also guide our efforts to ensure the sustainability of the framework of tools and guidance with which we would encourage the enhancement of research by using metadata actively. By exploiting the benefits of recording and sharing process and the underlying decisions, the framework would also enable one readily to use metadata actively to enhance research in the same way as other researchers from a range of disciplines would have done before.

Our longer-term aspirations would be to provide a way to compare metadata from different processes and disciplines, and potentially to provide a 'service' whereby anyone could generate a model for their particular metadata. New models could then be compared to those already in the framework and in principle the service could suggest similar data/metadata architectures. This strategy would obviate any pressure to become an expert metadata architect to understand and implement the concept of active use.

We have already demonstrated the power and potential of using metadata actively in large-scale experimentation, climate modelling, tomography, and in sound arts practice. We would actively investigate other application areas, including crystallography and synthetic chemistry.

In concluding this paper, we invite all interested parties to consider how they might adopt the active use of metadata in the furtherance of their own research.

## Acknowledgements

The authors are grateful to Jisc for bringing them together as part of the Research Data Spring project, enabling them to coalesce their common interests in enhancing research by making better use of metadata under the auspices of the CREAM project, Grant References 3544 and 3742.

## References

De Roure, D. (2010, November). *Replacing the paper: The twelve Rs of the e-research record.* SciLogs. Retrieved from http://www.scilogs.com/eresearch/replacing-the-paper-the-twelve-rs-of-the-e-research-record/

Fauconnier, G. & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. New York, NY: Basic Books.

Garrelfs, I. (2015). Procedural Blending. Retrieved from https://blog.soton.ac.uk/cream/discussing-active-metadata/%20procedural-blending/

Garrelfs, I. (2016). *From conceptual blending to procedural blending: Applying a model of cognition to process in sound art practice*. In: Denham, S. and Punt, M. eds., (2016). Off the Lip Conference Proceedings 2015. Plymouth: TT OA Papers.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., & Kleese, K. (2010). Using a core scientific metadata model in large-scale facilities. *International Journal of Digital Curation 5*(1). doi:10.2218/ijdc.v5i1.146

Milsted AJ, Hale JR, Frey JG, Neylon C (2013) LabTrove: A Lightweight, Web Based, Laboratory "Blog" as a Route towards a Marked Up Record of Work in a Bioscience Research Laboratory. PLoS ONE 8(7): e67460. doi:10.1371/journal.pone.0067460