Towards Transparency in Visualisation Based Research

Drew Baker King's Visualisation Lab,
The Centre for Computing in the Humanities.
King's College London, UK.

Abstract

The composer Felix Mendelssohn said "music is not too indefinite for words, but too definite." Similarly 'data objects', objects about which data is held, have often been considered to be too vague and have been pinned down to specific and objective categories through the use of *metadata*[1]. This paper examines data and its metamorphosis and proposes that there exists a parallel stream of ancillary information to *metadata* which is generated as part of a visualisation-based research process, and which it is necessary to document and disseminate alongside the visual research outcomes.

This "*paradata*", the paper argues, is essential to understanding and building successful and transparent research hypotheses and conclusions, particularly in areas where data is questionable, incomplete or conflicting, and explores how this can be applied to the process of creating three dimensional computer visualisation for research.

Definitions

This paper will use a number of terms which are defined here.

Data Object
A 'thing', perceptible by one or more of the senses, about which data is held.

Data Artefact
Data recorded and attributed to a data object as a product of human conception or agency.

Data Load
The amount of data artefacts created and assigned to a data object. A low data load suggests little is recorded about the data object*;* conversely a high data load suggests that there is a large number of data artefacts associated with it.

Metadata
A data artefact attributed to a data object describing an inherent element of the object.

Meta-data
The schema and structure of data used to describe metadata.

Paradata

[1] **Plagman, B 1974** "*Data Dictionary/Directory Systems*", Wiley.

As will be shown, a data artefact attributed to a data object (or data artefact) describing a subjective element of the object.

Introduction

It is almost universally acknowledged that the inclusion of metadata in information management and recording is, if not vital, extremely important in understanding data. Metadata is often described as 'data about data' (or in the case of *meta-data,* data about data elements or attributes, e.g. height, length, weight, time of creation etc.); as such it attempts to describe information in a more comprehensive way, providing a deeper understanding of the information behind or at a higher level.

Amongst the various metadata schemes, the most widely accepted metadata in the Humanities (though the lens of Digital Culture) is the Dublin Core Metadata Element Set[2], a base of fifteen fundamental 'facts' that may be attributed to a digital object: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

Each of these may be repeated or omitted and each has defined schemes to help clarify the description. Within Dublin Core, for instance, the 'Type' element has twelve recommended terms: collection, dataset, event, image, interactive resource, service, software, sound, text, physical object, still image, moving image.

This approach facilitates the precise description of data objects, in many cases using controlled vocabularies to ensure as broad a consensus as possible of understanding how the metadata recorded is to be understood (e.g. agreed formats for entering place-names, dates etc.). This is ideal if we wish to systemise the recording of information regarding data objects. However, it is the proposition of this paper that another form of '*metadata*' exists – '*paradata*' – that does not directly describe the data object*,* but rather lies alongside it, describing the dynamic nature of the process involved in interpreting and creating visual representations of data objects.

Physical or quantifiable properties of data objects tend to be fixed (e.g. the height, weight, find-location etc. of an object) and ideal for categorising using the appropriate metadata schema to create a data artefact of the data object. The properties of the discussions, interpretations and decisions that constitute the process of creating a visualisation of an object, which may not physically exist, are also valuable data artefacts but by their ephemeral nature are more difficult to define.

While both metadata and paradata are similar, the aims of each are subtly different. Metadata tends to describe more or less static information about a data object: paradata describes the process of interpretation of a data object. Metadata records the documenter's interpretation of a data object: paradata records the documenter's *process of interpretation*, to allow it to be subjected to scrutiny and evaluation.

---

[2] Dublin Core publication details online http:/www.dublincore.org

This assertion of the existence of paradata presents the humanities research community with a treasure house of potential to be exploited if method(s) can be successfully devised to:-

1) Identify the decision-making processes that occur during visual research.
2) Capture the intellectual capital invested within that research.
3) Record paradata with the minimum overhead within a project.
4) Disseminate the paradata in an easily accessible form.

The extent of this task is broad and beyond the scope of any single paper. This paper will therefore discuss the fundamental tenant of paradata, data metamorphosis and argue the case for paradata capture within visual based research outcomes.

Data Metamorphosis

Key to solving this challenge is the identification of the initial point at which data changes, or metamorphoses, through the process of combination with other data objects and becomes transformed into useful information.

Data by its very definition has no intrinsic value; it exists within its own right but is valueless unless it is placed within some context. When data (e.g. '9.10') and meta-data (e.g. 'Height in cm') are combined, some property of something is described (e.g. 'Height in cm = 9.10'). This is information, but of limited value, as its significance is not known. When some sets of these descriptions are combined, more meaningful information is produced, e.g. 'Name = Vase 7014b Height in cms= 9.10'.

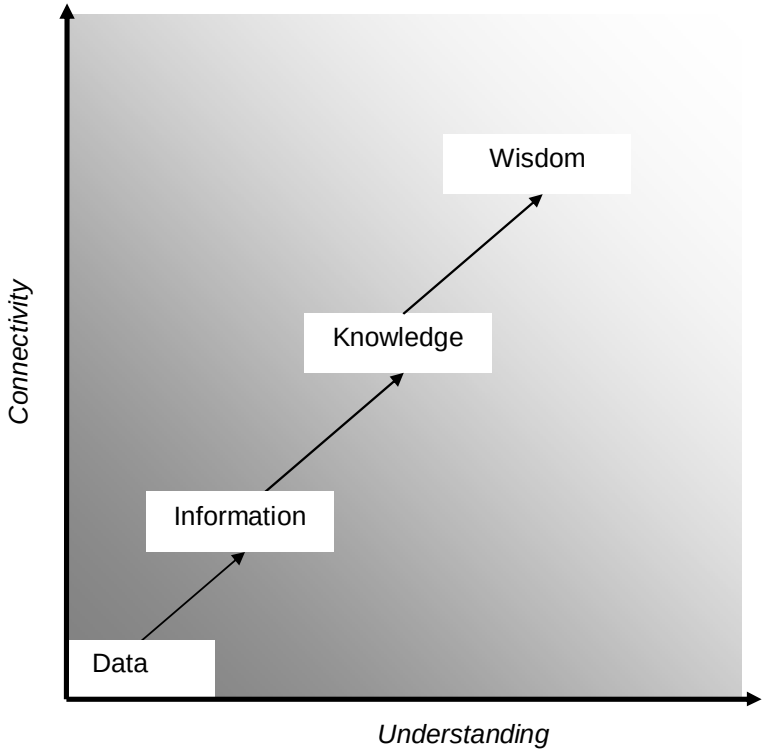
A combination of such descriptions produces a record or 'data artefact' about an object. Note that the data themselves do not change, but rather their significance changes as they are combined with other data and metadata, increasing the 'data load', the quantity and extent of information that the record or data artefact contains.

Metadata is a means of recording data about a data object and its associated data artefacts, and relationships between data, in a meaningful way. However, metadata is less fit for describing the analytical or interpretative processes that researchers bring to bear upon artefacts, which it is essential to understand if data artefacts are to be validated.

Visualisation-based research both draws upon data artefacts as evidence, and creates both new data objects and data artefacts (the visualisations and records of them). The idea of paradata represents an attempt to address the critical absence of a means of adequately describing researchers' analysis and interpretation of both evidential and visualisation objects and their associated data artefacts.

One of the popular models of data metamorphosis is the "Data, Information, Knowledge, Wisdom model" (DIKW) proposed by Cleaveland, Zeleny, Ackoff [3] and others. This model is used by both the disciplines of information science and knowledge management to create a hierarchy of data 'worth' as a data artefact moves though stages of transformation. The model's proposition is that data transformation occurs as understanding increases and data artefacts become more connected.

3 ***Ackoff, R. L**., "From Data to Wisdom", Journal of Applies Systems Analysis, Volume 16, 1989 p 3-9.*



The dilemma of the DIKW metamorphosis of data is that the further data is abstracted from its original source through interconnection with other data elements, the more prone it is to errors entering the metamorphic chain. Each contextualisation, assumption and decision made on data (and its products), while potential increasing the understanding of the data, also makes it potentially less ‘pure’.

Further complexities are involved when additional data artefacts are brought to bear in the linier metamorphic process; where do they fit in, what are the relationships between data artefacts, and what are the processes that form or lead the decisions made during the visualisation process?

It is not the purpose of this paper to criticize an established data model but rather to use its principles to show how visualisation research is currently used and presented. While the visualisation itself is grounded in ‘data’, factual or hypothetical, and may even be used to construct the narrative thread of the accompanying textural work, the processes and designs of the construction of the work are opaque to anyone outside the originator unless specifically referenced.

The final result, usually a set of visualisations presented as part of a publication (often supporting the main body of work as illustrative ‘proof’), apparently emerges fully formed and without any contextualisation - much like ‘wisdom’, it has to be reinterpreted by the recipient. This is where many misunderstandings, misappropriations and critical attack are propagated, simply because the underlying data, processes and context are absent from the visual image.

The Need for Paradata

If we can say that metadata is the information which is used to describe a data object, then is it possible to use metadata to track the metamorphosis of the data artefacts associated with it without ascribing a new 'buzz word'?

Although similar in kind (both are forms of 'data'), metadata and paradata are different in type (quality), not just in degree (quantity).The crucial distinction is that, while metadata pertains to the data object, *paradata* pertains to the process of analysis and interpretation of the object and its data artefacts. Paradata is chronologically and informationally dependent upon data, but paradata does not constitute a type of 'hyper' or 'superdata' pertaining to *data artefacts*; rather it pertains to the critical process. Hence, for paradata, terms such as interpretation, knowledge, understanding, synthesis etc. are key concepts that metadata does not capture.

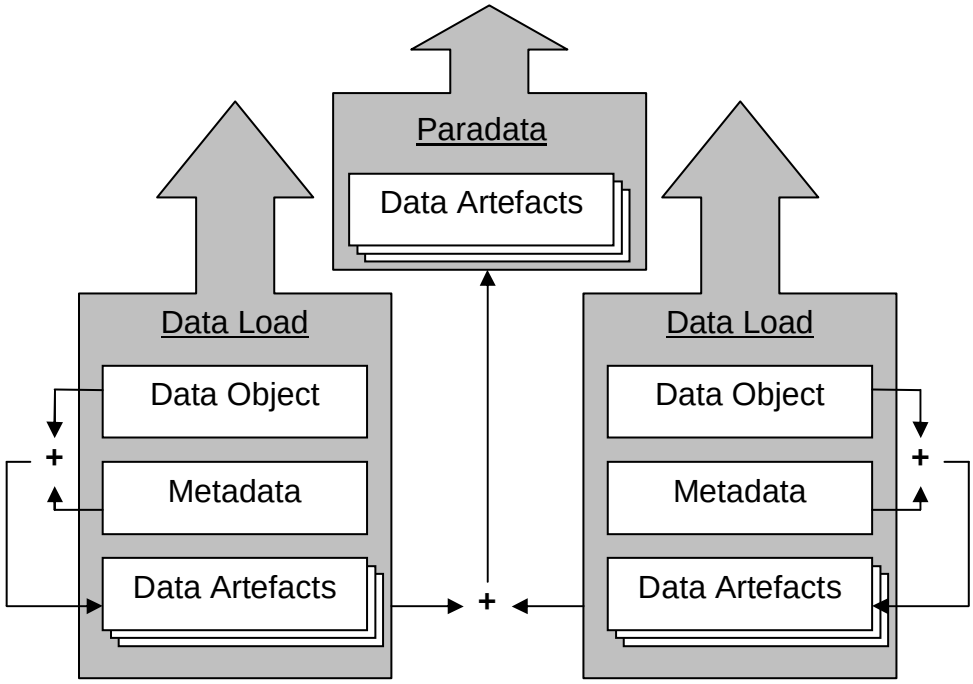
This has resonance with Polanyi's[4] dimension tacit knowledge "Into every act of knowing there enters a passionate contribution of the person knowing what is being known and that this coefficient is no mere imperfection but a vital component of this knowledge". This subjective 'imperfection' can not be data as it comes into existence each time data is transformed, neither can it be metadata which strives to be objective. Rather it is a data stream which contains the decisions, selection processes and reasoning behind the interaction and combination of different data artefacts. This data exists parallel to the data object and associated metadata artefacts.

As metadata artefacts are refined they are added to the data load for their parent data object. These historical records are not discarded or destroyed, and they become valuable to the researcher in understanding the objective context of the data object. By extending this to encompass the subjective paradata records of observations, decision and reasoning processes and selection and rejection criteria, similar contextual value can be added to the object's data load.

This subjective approach to the data object liberates the research by permitting the inclusion of clear reasoning behind not only the path the research takes but also the reasons behind omitting data through selection, distractions from the research narrative and even failures.

---

[4] **Polanyi, M. (1958).** *Personal Knowledge: "Towards a Post-Critical Philosophy*. University of Chicago Press, Chicago

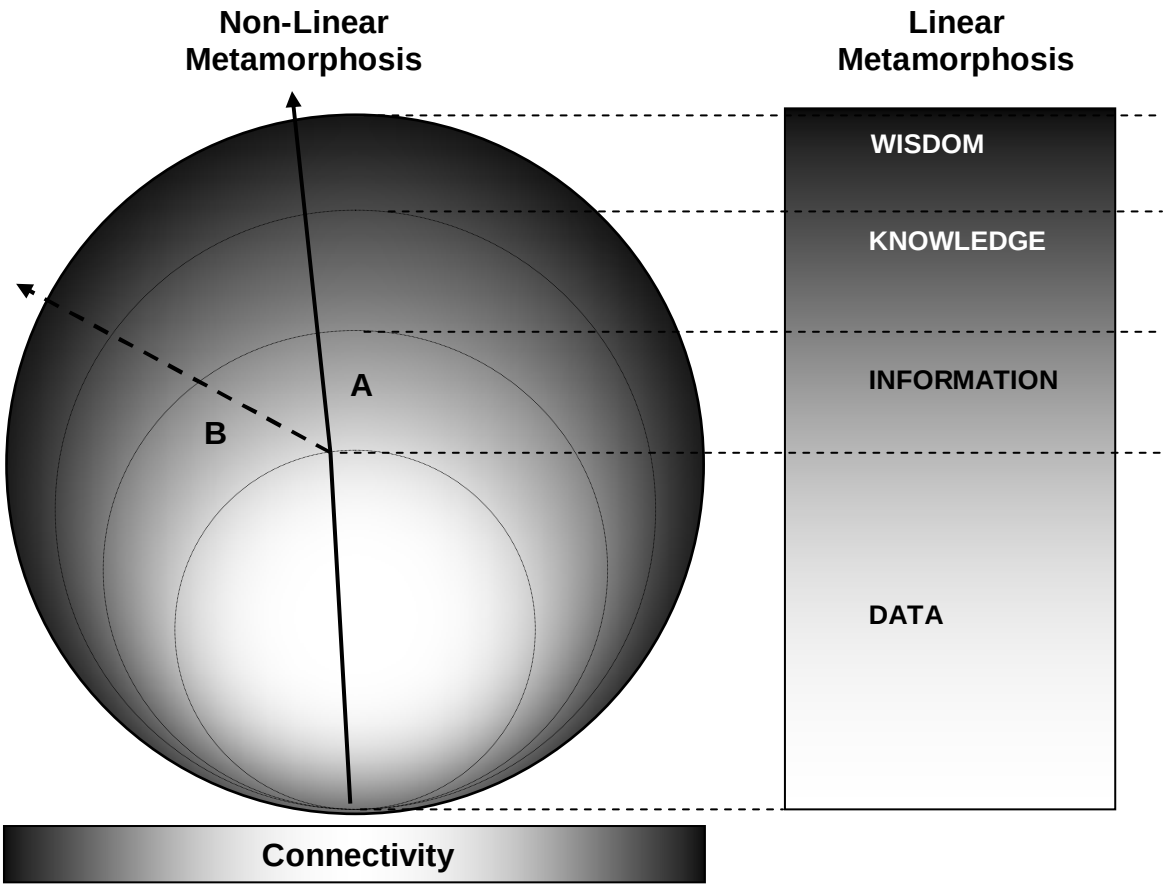


Separating the metadata load from the paradata load has the clear advantage of delineating objective observation from the subjective process of transformation. Paradata can therefore be said to contain such subjective things as selectivity, evaluation, exploration of ideas, entropy, cultural assumptions and research, stylistic decisions, inference and implied possibilities and probabilities.

By capturing this ephemeral data, it is possible to track the reasoning and construction of the visual hypotheses presented in a visualisation. The output of the visualisation process becomes more than simply a ‘pretty picture’ derived from the research narrative; rather it becomes *the* research narrative affording others the ability to track the argument as it is constructed and see the process and concepts behind the visualisation.

Instead of a linear narrative to a theoretical reconstruction, paradata allows the scholar to see the journey to the final visualisation. The discovery process is able to be more varied because the decisions and rationales employed by the researcher in developing the visualisation are recorded and more transparent; thus the linear degradation of ‘purity’ over understanding implied by the DIWK paradigm (as shown above) holds less true. Assumptions, beneficial wishful thinking, leaps of faith, place holder objects, arbitrary decisions and even whimsy can be introduced into the visualisation development cycle tagged by paradata and accessible to the end recipient.

Shown in the figure below, the point of metamorphosis, if such exists, now appears more like a start of a ripple on a pond; the further the ripple from the origin, the more potential for corruption or misunderstanding occurs. However, the transformations between the boundaries are recorded in the paradata and the decisions, logic, and reasoning behind the direction taken by the researcher are better comprehended.



The path of research A is clearly defined and digression from it onto a different path (B) builds upon the groundwork already set out. Because the paradata has been included within the project path, the decisions and process made can be inspected giving enhanced possibility for scholarly debate, inspiration and reuse of visualisation components.

The paradata records the processes the researcher has undertaken to develop the final visualisation. The journey now becomes as important as the destination itself.

Limitations

While the implementation of methods to acquire paradata and capture the intellectual capital behind a visualisation is clearly valuable to research, it is acknowledged that it is not without its limitations.

There is a significant concern regarding the granularity of the paradata that should be recorded. Clearly there are some major data items that must be recorded such as the evidence that informed the creator, the reasoning and decision made on that evidence and how it was implemented within the visualisation; but the finer the information recorded the, arguably, less valuable it is to the project itself.

Aligned to this is the question of the quantity of the paradata that should be recorded. The evidence upon which reasoning is based should be recorded but there will be instances, for example when using component parts, where the question of recording each individual component must be questioned (e.g. when a column is used, should the column be recorded or each of its components and subcomponents).

If we must consider granularity and quantity of paradata recorded, we must also consider its quality. How much weight do we attribute to a data artefact and how reliable is external paradata that may be anecdotal, biased, not directly relevant or “tainted” in some form but still potentially valuable to the research corpus as a whole?

Arguably the time taken to record paradata in addition to metadata and the actual production of the visualisation itself is the biggest single limitation of the paradata concept. A balance needs to be struck between the time taken to record paradata and the actual research to be conducted. While ultimately it is considered that the recording of paradata will be beneficial to the research process, its physical creation and recording not only takes time but to a degree stifles the creative thought process of the creator.

This is by no means an exhaustive list of the type of questions that need to be investigated and a successful compromise made if paradata is to be used effectively without suffocating the research objectives in another exercise.

To this end the London Charter organisation has adopted the principle tenants of paradata to develop and establish internationally recognised principles for the use of three dimensional visual research outcomes within the Arts and Humanities community.

Conclusion

It is the conclusion of this paper that a new form of data – paradata – does exist and that if this data stream can be tapped into, it will provide researchers with a better understanding of the process of creating visualisations. It will also provide scholars with a new way of debating and using data artefacts within visualisations and will reduce the use of computer graphic images for their own sake that may be misleading in their interpretation.

By understanding and marking the journey the researcher takes, the evidence which is used and discarded, the paths that are taken and where they lead to in the construction of the research narrative, the visualisation that is created becomes more than the pretty picture: it becomes a research artefact in its own right that provides the stimulus for debate and rhetoric, and it becomes a living growing data corpus.

However it is acknowledged that a number of activities must be conducted before we can proceed with the implementation of paradata within scholarly research.