Dataset Publishing - a Means to Motivate Metadata Entry

Shawn Callahan, Consultant, Strategic Management Sciences
B David Johnson, Principal Consultant, EarthWare Systems
Paul Shelley, Manager National Directory of Australian Resources, National Resource Information Centre

Introduction

Scientists are leaders in the use of metadata. They have assumed this leadership role through 2 key requirements (particularly over the past 10 years): (1) to effectively manage increasingly larger and complex datasets; and (2) working in interdisciplinary teams where data sharing is essential.

The volume of data has reached a level where scientific organisations must look to ways to effectively manage their dataset assets. There is a considerable risk of data loss or expensive re-acquisition due to lack of knowledge of data holdings. The ability to quickly find a dataset and determine its relevance are paramount requirements. A primary aim of a growing number of scientific organisations is to facilitate dataset re-use; thereby, in economic terms, providing a greater return on their investment. To date, however, very few dataset catalogues have delivered the expected benefits.

From the outset, metadatabase developers have focussed on defining (a) what should be recorded about any given dataset (eg. the Federal Geographic Data Committee's *Content Standards for Digital Geospatial Metadata*, 1994); and (b) the database application that will store and manage the metadata (FINDAR, GENIE, NASA Master Directory). While the above are critical components in the development of a metadatabase system, metadata administrators consistently express concerns over the difficulty of obtaining and maintaining the catalogue information.

Motivating people to document datasets is a key challenge. Scientific datasets require much interpretation that can only be provided by a human, and some aspects (being inherently subjective) will never be automated (Strebel, *et al*. 1994:2). Documenting datasets, however, is often viewed by dataset authors as an onerous task and is therefore left undone.

In an ideal world, datasets would be self-documenting and tools would be available to automatically extract the required metadata from a dataset. In reality, however, a considerable proportion of a given dataset's description is stored in the dataset author's head, and the successful extraction of quality metadata is therefore dependent upon that individual's full cooperation. It follows that individuals, not organisations, determine the time and effort expended on the task of dataset documentation. Consequently, we must redirect our attention from the technical solutions (eg. software development), to finding innovative ways of motivating dataset authors in the first instance to actually document their datasets, and in the second, maintain that documentation.

The authors have been involved in the development of numerous scientific dataset catalogues over the past 10 years (Johnson, *et. al.*, 1995; Callahan, *et. al.*, 1995; Johnson, *et. al.*, 1991) and during that time we have come to realise that information technology is no longer the impediment to developing successful metadatabase systems. It is our view that the key to successful metadatabase implementation lies primarily with the people and organisational issues, particularly in providing sufficient motivation for individuals to document their datasets. This has lead us to further examine the dynamics of behaviour of individuals within organisations. Work undertaken by Fraser (1995) on the motivational aspects of information systems success led us to investigate behavioural theory.

Naylor, Pritchard and Ilgen's (1980) theory of individual behaviour in organisations is applicable to the issue of motivating dataset authors to effectively undertake the task of dataset documentation. Naylor *et. al.* (1980:2) point out that the factors influencing behaviour are complex, and are a "composite ... of a number of different psychological 'processes,' such as learning, motivation, and perception." Their theory of individual behaviour in organisations, known as NPI Theory, attempts

to address why individuals undertake certain acts at a given level of commitment, in preference to other acts with alternative levels of commitment. This paper uses the key points of NPI Theory as the basis for developing a method for motivating potential dataset describers to participate in metadata population and maintenance.

The authors have built upon work conducted on behalf of the Australian Geological Survey Organisation (AGSO) and expand the ideas presented by Callahan and Johnson (1995). We have presented our discussion in three parts: (1) a discussion of NPI Theory as it applies to metadata population and maintenance; (2) a description of the AGSO solution, and how it relates to NPI Theory; and (3) general guidelines for metadata population and maintenance.

Although this exercise focuses upon developing a metadatabase that describes scientific datasets, the experience may be more generally applied to the rapidly emerging developments in data warehousing. Inman and Hackathorn (1994) recognise metadata as a key component of a data warehouse. The lessons learnt in the scientific arena are easily translated to these new developments.

Motivation and NPI Theory

NPI Theory describes the factors which influence an individual's behaviour within an organisation. More specifically, the theory deals with "why the individual chooses certain alternative courses of action in preference to others, and thus it might properly be called a theory of choice behavior." (Naylor *et. al.*, 1980: 3). The purpose of studying NPI Theory in the context of metadatabase implementation is to gain an understanding of the various factors that could influence an individual in determining how much time and energy will be expended in documenting a dataset in the face of other competing needs. The component of NPI Theory most relevant here is motivation.

Naylor *et. al.* (1980: 159) define motivation as "the process of allocating personal resources in the form of time and energy to various acts in such a way that the anticipated affect resulting from these acts is maximized."

The amount of 'time and energy' devoted to a particular act is not usually exclusively allocated (the exception being in times of extreme danger, such as an immediate threat to life where all energy is devoted to the act of avoiding death). Naylor *et. al.*, (1980) point out that a person's job can be seen as a *pattern* of outputs that constantly compete with one another for the allocation of personal resources.

The choice of which acts gain priority is initially governed by an individual's *needs*. The needs from which one is deprived tend to provide the greatest motivational force. NPI Theory states that the anticipated affect (perceived results based on performing particular acts) "is partially determined by the current level of need deprivation." (Naylor *et. al.*, 1980: 161). The higher the deprivation, the greater an individual's motivation to satisfy that need.

The meeting of needs as a source of motivation

The motivational force is driven by an individual's *basic needs*. Murray (1938: 80-83) lists what he terms the primary and secondary needs of individuals. The former are characterised by innate biological requirements, such as the need for air, water, food and sex. These primary needs tend to have little effect on the time and energy an individual expends documenting a dataset.

The secondary needs have no localisable bodily origins and pertain to learned or acquired psychological factors. Murray (1938: 80-83) lists twenty eight needs of which many are potentially relevant to metadata population, including the need for acquisition (eg. monetary reward), achievement, recognition, exhibition, autonomy, affiliation, order, cognisance and exposition.

Needs are the motivational well-spring, but they are dynamically filtered and altered depending on the many complex factors that comprise the decision making process. Consequently all needs, at any particular time, vary in their motivational affect.

Anticipated affect as a motivational force

"It is only through understanding the person's expectations of the future in terms of perceived contingencies, outcomes, and anticipated affect associated with these outcomes that we can really explain motivation." (Naylor *et. al.*, 1980: 160).

The term 'anticipated affect' indicates that motivation is future-oriented. The decision to direct resources to a given act is strongly influenced by knowledge of the outcome, where the individual is most likely to choose to undertake acts which are seen as having a positive (maximising) outcome, and avoid those which are seen as negative. The envisioned anticipated affect is based on: past experience; information provided from a credible source; or through observation.

The anticipated affect is explained through an interaction of four components: act, product, evaluation and outcome. In NPI Theory, the interaction between these four components is termed *contingencies*, whereby the components are coupled to form defined relationships such as: (1) Act → Product; (2) Product → Evaluation; and (3) Evaluation → Outcome. These contingencies will describe the individual's perceived relationship between the three component pairs; what product can an individual expect to produce by undertaking a particular act; what evaluation can they expect for producing that product; and what outcomes are likely if a particular evaluation is provided?

Act → Product

In terms of motivation, the ideal situation is one in which greater levels of time and effort will create a commensurate quality or quantity of product. A de-motivating situation would be one in which there is little or no increase in product quality or quantity no matter how much effort or time is expended.

At this stage there is no evaluation of the product's worth, but the activity to product ratio must be in accordance with the individual's expectations.

Product → Evaluation

"The salience of products to the individual is a function of (a) whether these products are evaluated; and (b) the power of the evaluator to make valued rewards contingent on that evaluation." (Naylor *et. al.*, 1980: 43)

An unevaluated product is worthless, because no worth has been consciously attached to it. An evaluation can be made by the individual creating the product, or by some other person.

It is also important for the individual to know that the evaluator has the capability to influence the outcome (eg. they have the decision-making power to promote the individual within the organisation).

To effectively evaluate a product, a system of criteria is required, which will, for example, define terms such as quantity and quality If the individual understands those criteria, and therefore can judge the likelihood of a favourable evaluation, they will be in a better position to assess how much time and energy they need to commit to the creation of a given product.

Evaluation → Outcome

The outcome is the final component in the chain, and is probably the most important in determining the motivational force. It is here that the perceived benefits are realised. It is the outcome that potentially satisfies the individual's basic needs.

The outcomes must be clearly defined, and be based upon a criteria system derived from the evaluations. The criteria system must be applied consistently over time.

Contingencies and metadata population and maintenance

The above discussion highlights the importance of each step in the contingency chain. If any point in the chain is weakened, the potential motivational force is considerably reduced. It is important, therefore, to develop strategies that take account of each link in the chain, and their interrelationships.

Figure 1 illustrates the ideal (ie. high motivational force) contingency relationships as they relate to metadata population and maintenance.

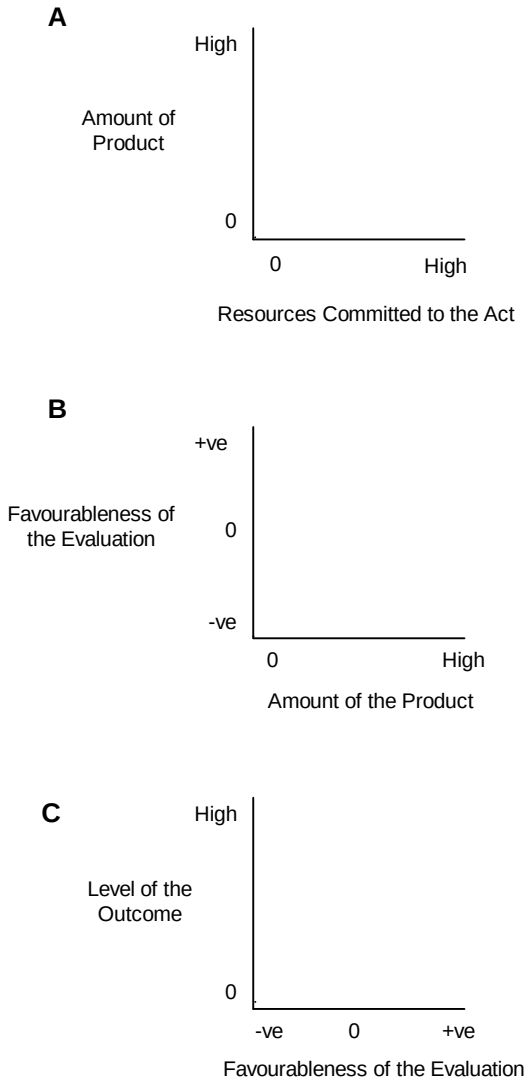

Figure 1: Ideal Contingency Relationships (after Naylor et. al., 1980: 174)

Example A illustrates that wherever possible, the resources required to undertake the documentation remain relatively low compared to the amount and quality of product generated. This can be achieved through appropriate use of user-friendly applications, tools designed to automate the metadata population process, and efficient procedures. As noted previously, however, there will always be a requirement for human intervention, but we should endeavour to minimise the effort where ever possible.

In Example B we see that if little or poor quality product is generated, a negative evaluation is bestowed (eg. if no datasets are described, the individual can expect a negative evaluation) until a pre-determined threshold is met, and increasingly favourable evaluations are achievable thereafter. A favourable evaluation will bring outcomes that satisfy needs such as recognition, career advancement and monetary reward.

The expected corollary follows then, that it is only when the favourableness of the evaluation is positive (example C) that outcomes start to accrue.

These graphs provide a visual depiction of the requirements that must be met to establish a successful metadata population and maintenance program. It can be seen that a central requirement of the program is borne from the fact that an individual will be motivated when they feel certain that "*different* levels of commitment to the act result in *different* levels of anticipated affect" (Naylor *et. al.,* 1980: 189).

The following section describes the solution provided for the Australian Geological Survey Organisation (AGSO) to develop and maintain a metadatabase to manage their extensive spatial dataset collection. In this brief case study we describe how NPI Theory applies to the AGSO solution. We were unaware of NPI Theory whilst developing the AGSO solution. Consequently this is an ideal opportunity to test the theory against the system we have proposed for AGSO.

Australian Geological Survey Organisation (AGSO)

The authors were commissioned by AGSO in 1995 to assist in the establishment of AGSO's new metadatabase system, and this project provided an ideal opportunity to demonstrate the importance and function of motivational issues in the establishment and maintenance of a metadatabase system.

The Australian Geological Survey Organisation is a primary collector of geoscientific information in Australia. AGSO required a system to enable clients (internal and external) to manage, find and interpret the many geoscientific datasets collected by AGSO. In recognition of the need to effectively manage AGSO's datasets, the authors were initially commissioned to develop a prototype metadata system for datasets which are to be widely used.

Following is a description of the authors' proposed solution.

Dataset publishing analogy

The concept of *dataset publishing* was developed to address the challenge of motivating people to provide and maintain dataset descriptions. AGSO's scientific culture currently rewards people for publishing scientific papers. These rewards are built into the merit promotion scheme and form part of an individual's performance appraisal goals. We proposed that creating a dataset was akin to publishing a scientific paper, thus providing dataset authors with an additional mechanism for gaining recognition. AGSO staff understood the concept of publishing and consequently could immediately see the benefits of the approach.

The publishing concept involved the establishment of procedures, the allocation of human resources to manage these procedures, and the development of software to assist in the collection, maintenance and dissemination of the dataset descriptions. The process of publishing a dataset is as follows.

The dataset must first pass through a series of steps similar to publishing a conventional scientific paper including, for example, adherence to notes for authors, peer review, and editorial refinements.

As a key requirement of publication, the author is required to document their datasets according to a set of standards (these standards have a similar function as notes for authors in a conventional paper). The outcome of this documentation is the acquisition of metadata which describes the published dataset. In NPI Theory terms, these standards define the criteria system for evaluation.

People who develop a dataset are known as *dataset authors* and are primarily responsible for the content of the dataset and dataset description. There are many levels of publication, reflecting the intellectual effort invested in its creation (for example, a fully reviewed scientific paper in a recognised international journal as compared with an internal discussion paper). It was recommended that AGSO determine where a published dataset fits within the current scale of publications. An important point derived from NPI Theory was the idea that varying levels of an act should result in varying levels of outcome. Our recommendation to AGSO to fit a published dataset (as a single item) into a scale of publications doesn't fully adhere to the requirement of having varying levels of dataset publication. In AGSO's case, there should also be a scale of dataset publication.

The *Dataset Editor* and *peer reviewers* determine which datasets are published (do they meet the standards?) and the publication is linked to specific outcomes. The Dataset Editor is responsible for ensuring dataset authors are adhering to the standards (cleaning data where appropriate), ensuring

These software applications reduce the effort required to undertake the act of producing a dataset description. In terms of NPI Theory, this is an advantageous solution, as the time and effort required to undertake the act is significantly reduced.

Ancillary benefits of the dataset publishing analogy

There are a number of ancillary benefits to the dataset publishing analogy. Firstly, the creator of the dataset becomes the dataset author, which immediately bestows the responsibility for the content of the dataset with an individual or group of individuals. Detailed questions about the dataset can therefore be directed to the dataset author, rather than to a publisher who may know little about the actual dataset itself.

Secondly, published datasets will have undergone a number of quality control processes and therefore have a known and demonstrated level of quality. The organisation can therefore feel confident with the stated veracity of the data and users will be provided with data that meets their expectations.

As more datasets are published, the organisation will gain further recognition for producing high quality geoscientific datasets. This improves the status of the organisation as a whole. AGSO is committed to providing a high level of customer service, and this approach will greatly assist them in achieving this objective.

Guidelines for Metadata Population and Maintenance

NPI Theory provides a sound framework for determining the requirements for motivating people to document datasets. In AGSO's case, dataset publishing was the mechanism implemented to provide sufficient motivational force. Other organisations will require alternative solutions, but the basic tenets remain true. Following are seven general guidelines for motivating metadata population and maintenance.

1. Determine the individuals' basic needs. In many cases those needs will include recognition, achievement, or monetary reward, but these and others will vary depending upon each particular individual and organisational culture.
2. Act → Product: Find ways to minimise the time and effort required to create the metadata entry. For example: provide tools that are easy to use; define a clear set of procedures; educate individuals on how to best apply these tools and procedures; ensure sufficient resources are available to undertake the task.
3. Product → Evaluation: Evaluations of the metadata entries must occur, against a clearly defined system of criteria. Ensure that the evaluator has the power to deliver expected outcomes. Metadata entries should be classified in a way that reflects the quality and amount of product created.
4. Evaluation → Outcome: Ensure that outcomes are commensurate with the varying levels of evaluation. These outcomes must be clearly established at the outset and consistently applied.
5. The implementation of the metadata population and maintenance program must be viewed as an ongoing, long-term process.
6. Given the dynamic nature of needs, the implementation program must have inherent flexibility to change with changing needs.
7. The above guidelines are interrelated, and Guidelines 1 to 4 must be addressed together in a single program, and omitting one will severely weaken the potential motivational force derived from that program.

Conclusion

Based upon our experience in implementing metadatabase systems, the authors strongly believe that issues of motivation will significantly influence the success or otherwise of metadatabase implementation. NPI Theory provides a useful framework for understanding the factors which influence whether an individual will be motivated to provide and maintain a metadata entry.

To date, many metadatabase systems have been strong in ensuring that the Act → Product relationship is met at least in part (ie. concentration on the technology and tools to support metadata entry, storage and retrieval), but in the area of procedure definition are often lacking.

Furthermore, the Product → Evaluation and Evaluation → Outcome contingencies have tended to be overlooked altogether, which has been a major failing in metadatabase implementation to date. There is a critical need to ensure that products are adequately evaluated against a clearly defined system of criteria, and that outcomes (usually in the form of some kind of reward) are consistently applied. If these contingencies are met, a successful implementation is far more likely.

In the case of the AGSO metadata implementation, the authors found that the mechanism of dataset publishing satisfied these motivational requirements. By taking into account the existing motivational culture (ie. publishing of scientific papers), the authors applied this already-accepted and successful mechanism of scientific publishing to the area of dataset development and documentation. This approach was readily accepted by the individuals within the organisation, because the Product → Evaluation and Evaluation → Outcome contingencies were accounted for, and in this case, already well-established and understood.

In the AGSO model, we addressed each of the three contingencies, however we did not place great enough emphasis on the importance of the power of the evaluator to provide the expected outcomes after evaluation. There is also a need to develop a scale of dataset descriptions, ranging from one which meets a set of minimum standards, to one which meets or exceeds all criteria.

NPI Theory is, of course, much more comprehensive and complex than our description here. We have simply extracted what we believe are some particularly relevant components of the theory as they may apply to metadatabase implementation. There are other components worthy of further investigation, such as the impact of roles within organisations, and the influence that the judgement process has upon individuals' behaviour.

Bibliography

Callahan, S.D. and Johnson, B.D. (1995) Scientific Dataset Catalogues. *Proceedings of Second AGSO Forum on GIS in the Geosciences*, Canberra, ACT, 29 - 31 March 1995.

Fraser, S. (1995) *A motivational view of information systems success: a reinterpretation of Delone and McLean's model*. Research Paper Series, University of Melbourne, Department of Accounting and Finance, Parkville, Victoria.

Inman, W.H. and Hackathorn, R.D. (1994) *Using the Data Warehouse*, John Wiley and Sons, New York.

Johnson, B.D., Callahan, S.D., Shelley, E.P., & Wood, G. (1995) Scientific Dataset Catalogues - People and Organisational Issues. Paper presented at the *Third National Conference on the Management of Geoscience Information and Data, organised by the Australian Mineral Foundation*, Adelaide 18-20 July 1995.

Johnson, B.D., Shelley, E.P., Taylor, M.M. and Callahan, S.D. (1991) "The FINDAR directory system: a meta-model for meta-data" in D. Medyckyj-Scott *et. al.*(eds) *Metadata in the Geosciences*, Group D Publications, Loughborough, UK.

Murray, H.A., (1938) *Explorations in Personality*. New York, Oxford University Press.

Naylor, J.C., Pritchard, R.D., and Ilgen, D.R., (1980) *A theory of behavior in organizations*. New York, Academic Press.

Shelley, E.P., Johnson B.D. & Callahan S.D. (1995) The National Directory of Australian Resources: past, present and future. Paper presented at *GEOINFO V, the 5th International Conference on Geoscience Information*, Prague, June 1995.

Strebel, D.E., Meeson, B.W. & Frithsen, J.B. (1992) 'Metadata standards and concepts for inter-disciplinary scientific data systems - II,' in *Proceedings of Scientific Data Management Workshop*, IEEE, 1992.

Acknowledgments

We would like to thank Peter Fox, Geoff Wood and Brent Heuer for their comments and inputs to this paper.