Data authenticity and data value in policy-driven digital collections

V. Bunakov, B. Matthews, C. Jones, M.D. Wilson
Science and Technology Facilities Council, United Kingdom

This is a postprint of the article published under "green" Open Access conditions in
OCLC Systems & Services: International digital library perspectives, Vol. 30 Iss. 4, 212-231.
doi: 10.1108/OCLC-07-2013-0025

Introduction

Digital archives are increasingly being required to manage a wide range of collections of digital artifacts, moving beyond traditional documents and images to a diverse range of media types including sound, video, research data, websites, and even software artifacts. These collections raise a range of new issues in their management. In particular, as the volume of content increases dramatically, the diversity and growth rate of large digital collections becomes too great for in-depth human management intervention to be either possible or economically viable; this is particularly acute in those research data collections where the scale of data acquisition has been increasing exponentially.

Consequently, there is a need for the preservation infrastructure to be automatically controlled by clearly-defined retention and disposal policies. The purpose of actively managing a digital collection is to ensure that it is useful for the user community the archive or library serves, a key aspect of this is that the collection being provided has value to the consumer. Collection acquisition, retention and disposal policies balance the potential value of the material against the costs of managing and preserving the collection to optimize the service provided to the customer. For a traditional physical library collection aimed at active users, the value of specific material will change over time, especially in factual disciplines where the state of knowledge moves with new discoveries and items can become no longer relevant to active researchers, but become part of the history of the subject. Time also has the effect of creating rarity, another aspect of value. Being able to quantify "value" for a specific collection in a more automated fashion will enable digital preservation infrastructures to work more effectively on the large volumes of content.

The SCAPE[1], ENSURE[2] and SCIDIP[3] projects have demonstrated designs and tested implementations of scalable policy-driven digital preservation architectures; however, modeling the value of digital artifacts collections requires more analysis, despite the presence of mature cost-benefits analysis frameworks (see Beagrie *et al.*, 2008 and Beagrie *et al.*, 2010). The ongoing move of established collections towards open data repositories (Griffin *et al.*, 2012) as well as application of the best management practices in digital curation, requires a conceptual framework for the more explicit notion of *digital collection value* that will be suitable to underpin a manageable and interpretable preservation policy.

In the business world, the value of digital collections can be modeled, for example, as the costs of not having preserved the items, or the potential revenue that the preserved items could yield. In the case of research data, the revenue is rarely an immediate goal; there are examples when data collected hundreds of years ago for reasons that are irrelevant now can nevertheless prove invaluable for modern science, e.g. for predicting climate change (Kiefer and Wilson, 2012). Research data is therefore a good example when proper modeling of data value requires other considerations apart from revenue expected or costs involved.

The problem of modeling data value is getting more acute with the advent of "Big Data" for which the natural sciences and social research are prominent but not the only sources. What to select for preservation; what data aspects and properties to retain throughout the preservation lifecycle; how reasonable aggregations and collections of digital content can be created which are suitable for the intended purpose: all these questions make data value considerations important for Big Data. One cannot rely solely on human judgment to quantify value; harnessing automated techniques underpinned by value models will be needed.

Methodology

As the notion of data value may imply different interpretations, one needs to define this concept to make it operable. In this paper, the authors conduct a top-down analysis of digital preservation domain from the IT Service Management perspective, and will show that a well-known concept of authenticity can be a natural candidate to underpin data value. The authors then perform a bottom-up analysis of how authenticity has been understood in digital preservation projects and reference models, and will show that authenticity allows a generalization not exclusively related to the topic of data provenance but may be coupled with the notion of data value. Once top-down and bottom-up analysis have met, the authors apply the newly-acquired understanding of data value to the concerns of managing data collections.

In this paper, the authors heavily reference the Open Archival Information System model (OAIS, 2012) and use OAIS concepts and terminology where applicable. They also introduce other concepts and terms that are mostly generalizations of those suggested by OAIS: as an example, a "digital preservation solution" is a generalization of "digital archive."

As the study should eventually facilitate the execution of preservation projects, it is worth clarifying its position in the preservation project lifecycle. A preservation project will typically go through a number of phases. (OAIS, 2012) concentrates on the *modeling* phase of the preservation project lifecycle, breaking down the preservation system into its major components and functions. (Conway *et al.,* 2011) focuses on strategic analysis, modeling and implementation. (Rothenberg and Bikson, 1999) considers all stages from the analysis through modeling and design to implementation considerations. This paper is focused on *strategic analysis* as an essential part of a larger preservation project lifecycle, with some suggestions for modeling.

Digital preservation as a service

Digital preservation projects are typically focused on the detailed analysis, design and implementation of *preservation solution* to provide the infrastructure for digital archives; however, what the potential users, or to use the OAIS terminology, Designated Community and other consumers[4] are actually interested in, is a *preservation service* with a preservation solution sitting at the core of the service and accompanied by other human- and machine-enabled components that enable the service and users objectives to be achieved. Possible components of a digital preservation service are shown in the Figure 1.

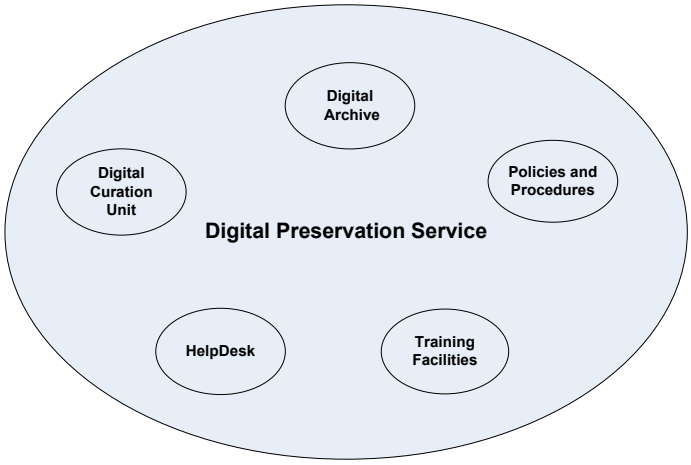

Figure 1. Possible components of digital preservation service.

There are always elements of service surrounding a preservation solution. As an example, even if a bare preservation solution is delivered in the form of a software application, then its users have to be supplied with the URL or installation instructions, some information about what the solution purports to be, and who is the contact for further enquiries. The authors do not focus on the service perspective in this paper, this is just a departing point, but it is beneficial to keep this perspective in mind throughout the entire lifecycle of a preservation solution analysis, modeling, design, and implementation.

If one is delivering a *preservation service* in digital preservation projects, then it is natural to consider digital preservation from the point of view of such an established discipline as IT Service Management and apply its concepts, terminology and practices. One immediate observation in applying IT Service Management perspectives to digital preservation is that the *preservation service objective* should be specified.[5]

ITIL, which is one of the prominent service management frameworks worldwide,[6] is concerned about *business value* that a service creates for its customers. The applicability of the business value concept to cultural and research domains that are the focus of digital preservation may be subject to discussion. Nevertheless, the importance of the customer perspective is reflected in OAIS that emphasizes the roles of Consumer and of Designated Community. Economic factors such as expected savings owing to the re-usability of digital information (in place of its expensive re-generation) or the transformation to another, less storage hungry format can be the drivers for some preservation initiatives.

Overall, the concept and the term “business value” is applicable to preservation projects and services even in a fiscal sense, but as the vision of business value delivered by digital preservation services is not limited to economic aspects, the authors prefer the more generic formula: the objective of digital preservation service, as a particular subclass of an IT service, is to provide *value* for its consumers.

Service Utility and service Warranty

In addition to a better understanding of the preservation service objective, the IT Service Management perspective can help to decide on what underpins this objective. ITIL suggests the concepts of *Utility* and *Warranty* as components of business value (ITIL Strategy, 2007). Utility is "functionality offered by a product or service to meet a particular need. Utility is often summarized as 'what it does'." Warranty is "[a] promise or guarantee that a product or service will meet its agreed requirements" and as "derived from the positive effect of being available when needed, in sufficient capacity, and dependably in terms of continuity and security."[7] Utility is what the customer receives, and warranty is how it is provided (ITIL V3, 2007).

The applicability of Utility to digital preservation seems to be clear, with the reservation that ”what it does” means preservation service in a broad sense including: the representation layer of a preservation solution in the spirit of the OAIS model; enabling information discovery and information retrieval capabilities, etc.

The applicability of Warranty to digital preservation seems less clear if taken literally: at first glance, it describes the non-functional capabilities of the service; however, if timescale is changed from short term to a long-term digital preservation perspective, then the meanings of "when needed" and "continuity" do shift accordingly, and Warranty is then applicable over the long and changing life of the preservation service.

Here is a schematic diagram for the concepts introduced or referenced so far using a non-formal notation:

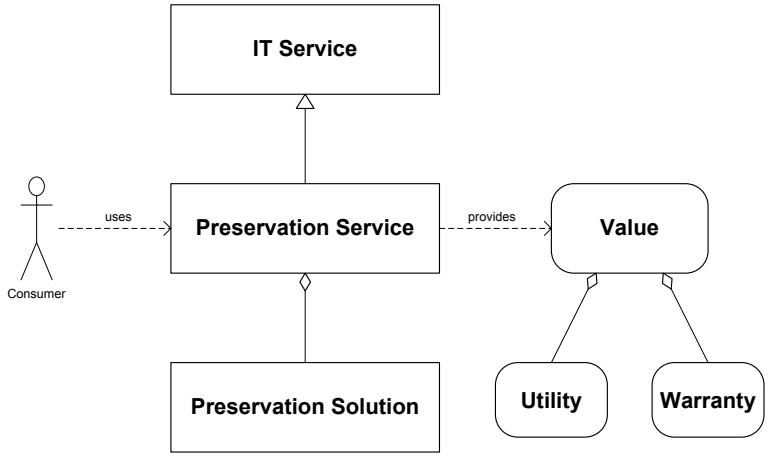

Figure 2: Breakdown of preservation service into conceptual entities

The ITIL concepts of Business Value, Utility and Warranty have something in common with the ISO 9000 family of standards related to quality management,[8] and with the traditional business concepts of "fitness for purpose" and "fitness for use."

- *Fitness for purpose* means meeting mission statements, proclaimed objectives, and stated outcomes of using a service or a system. In quality management terms, fitness for purpose is subject to *validation* which answers the question "Are we building the right thing?" An example of validation in an IT project is checking whether a specification captures all essential customer requirements (as well as adds no requirements taken just from analyst's mind), whether it is fit and sound as a foundation for the system design, and for the delivery of expected business outcomes.

- *Fitness for use* means the effectiveness of a design, manufacturing method, and support process employed in delivering a good system, or service that fits a customer's defined purpose, under anticipated or specified operational conditions.[9] In quality management domain, fitness for use is subject to *verification* that answers the question "Are we building it right?" An example of verification in an IT project is checking whether a piece of software actually implements the specification and does not produce negative side effects.

To make these considerations more explicit, see the Table 1 below:

	Utility	**Warranty**
ITIL definition	Functionality offered by a product or service to meet a particular need	Promise or guarantee that a product or service will meet its agreed requirements
Related business concept	Fitness for purpose	Fitness for use
Related quality management concept	Validation	Verification
Subject of checks	Customer needs and expectations	Requirements and specifications
Example of checks from IT practice	Whether specification captures all essential customer requirements	Whether a piece of software actually implements specification
Question "answered" by checks	"Are we building the right thing?"	"Are we building the thing right?"

Table 1: Utility and Warranty aspects of value in different disciplines and practices.

Authenticity from IT Service Management perspective

The IT service management perspective suggests that, from the consumer's point of view, the main expectation of any preservation service is its ability to support the information lifecycle in the consumer's interests. The authors think this is true not only for the Designated Community but for other types of consumers, too. Digital curators, librarians and archivists are interested in the information lifecycle for the same digital content but the stages in the lifecycle, how it functions, and what value it provides may be different from the Designated Community.

The information lifecycle is in essence the circulation of information entities that satisfy information needs of a particular consumer.[10] For consumer, having the ability to discover, retrieve, and handle the information to the information entity and having the means to validate that the obtained information entity is authentic (has all essential features and meaning) are the most important aspects of preservation service. One gets the information (that is provided by Accessibility), and one knows what one gets (that is provided the

Authenticity aspect); the combination of the two covers the consumer's information needs to full extent.

An information entity *manifests* its value for a consumer through the aspects of Accessibility and Authenticity. The focus on the two aspects of Access and Authenticity can be regarded as *a domain specific breakdown* of IT service value that reflects the needs and methods of digital preservation.

To conclude, Authenticity matters because it is one of the origins of *value* that preservation service brings to Designated Community and other types of consumers; the conceptual diagram is shown in Figure 3 **Access and Authenticity in another breakdown of preservation service.**

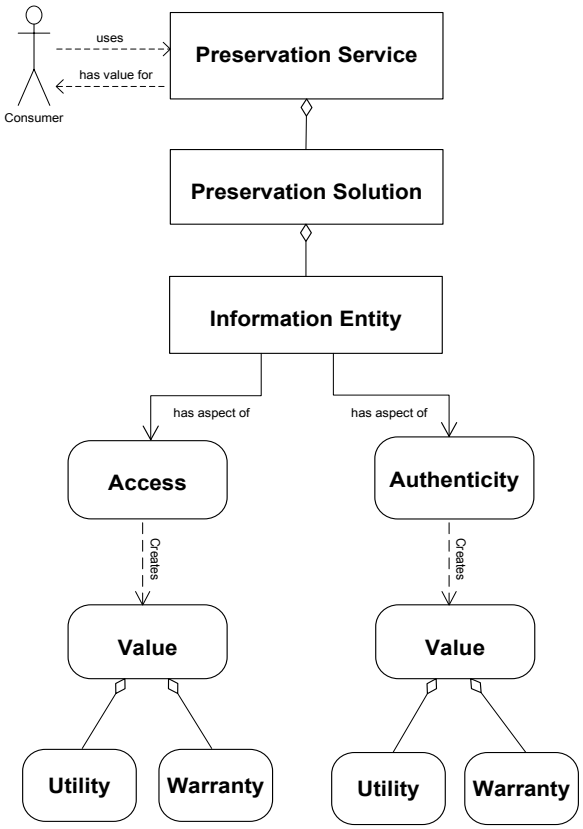

Figure 3 Access and Authenticity in another breakdown of preservation service.

This breakdown can be reconciled with the previous discussion on the preservation service objective. Usability and Warranty are distributed between Accessibility and Authenticity: Accessibility has aspects of Utility (e.g. letting consumers to reach the information) and Warranty (e.g. stability and security guarantees); a similar statement is true about Authenticity: it has aspects of both Utility and Warranty, which is discussed below. The aggregation of Utility and Warranty on the base level, and the sublimation of them while moving up to the top, contribute to the value of a preservation service as a whole.

This understanding of authenticity has been derived in a top-down manner from common principles of IT Service Management; in the following sections, the authors perform a bottom-up analysis to develop this vision, and then explore its implications for managing digital collections.

Related work on authenticity

The importance of proper interpretation of authenticity with its further application to the needs of digital preservation has long been in focus of researchers' interest: (Rothenberg and Bikson, 1999), (Rothenberg, 2000), (Lynch, 2000), (Levy, 2000), (InterPARES, 2002), (Guercio, 2008), (Guercio *et al.*, 2009), (Factor *et al.,* 2009), (Giaretta *et al.*, 2009), (Giaretta, 2011).

(InterPARES, 2002) selects *identity* and *integrity* as essential aspects of authenticity, and elaborates an analytical technique for determining the authenticity of archival records that rely on records *typization*. Overall, the (InterPARES, 2002) is a detailed exploration of the authenticity concept from the records management perspective.

The OAIS reference model (OAIS, 2012) defines authenticity as the degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence.

The OAIS-compliant authenticity model and authenticity management tool were in scope of the CASPAR project[11] that referred to results of InterPARES. In short, CASPAR modeled authenticity as an elaborated *process* tightly coupled with provenance management; if followed, the process gives the basis of evidence that is a part of the above definition. To manage the process, CASPAR designed a procedure called *Authenticity Protocol* consisting of *Authenticity Steps* detailed in (Guercio, 2008), (Factor et al., 2009) and (Guercio *et al.*, 2009). Further, OAIS indicates the importance of semantic representation information in order to underpin Significant Properties that in turn underpin authenticity (Giaretta *et al.*, 2009; Giaretta, 2011). This places authenticity in a subordinate role with its major application proving the integrity of the information transformations in on-going archive maintenance.

(Levy, 2000) provides insightful observations on what is the authenticity of *copies* including those resulting from transformations: the question is highly relevant in the digital era with the high plasticity of content. (Levy, 2000) rightly mentions that sometimes, like with minting coins or producing books on the printing press, there is no original at all, as all the coins or printed books are more or less copies of each other. (Levy, 2000) and also (Wilson, 2008) go on to observe that what is usually considered a digital artifact is in fact a *performance*, not unlike a performance of a play based on a script, or playing a piece of music based on a score. What is common between theatre or music performance and a book or a coin is that there is no *original* for them, yet one may consider the produced copies and performances quite authentic. Even the *matrix* from which performances are produced may not be precise or unambiguously defined. As an example, what can be considered “authentic Bach” may rely on his score that intentionally gave some freedom to the performers allowing them to improvise.

The same idea of performance but with references to the IT concepts is expressed in (Rothenberg and Bikson, 1999): virtually every digital artifact, even a plain text file, is in fact a *program* that has to be interpreted or executed in order to be consumed; e.g. the text file in ASCII format can be rendered by a printer and result in a paper artifact. This notion of performance has been extended to include preserving software packages in (Matthews *et al.*, 2010), which takes a view that authenticity of software lies in the extent it preserves the execution behaviour of the original in a new environment. (Pugh, 2006) expresses the same basic idea in different terms: he again takes a printer and a file in PostScript format, and states that printer *implements an interface* prescribed by PostScript. "The set of printers that understand PostScript can be considered polymorphic implementations of the PostScript interface," and if one wants to display PostScript on your monitor, you can use a viewer like GSView which is another implementation of the same interface.

(Lynch, 2000) takes a look at *trust* as an aspect of authenticity. This work as well as (Levy, 2000) pays attention to the fact that authenticity aspects as identity and integrity depend on the chain of trust that is conditional and subject to social influence. One thinks a certain *objet d’art* is produced by a famous master because the experts say so, but the experts themselves rely on a trail of evidence that in the end is rooted in a socially accepted agreement on the artwork’s origination.

(Levy, 2000) referring to (Smith, 1996) raises an important question of digital objects *boundaries* and *stability* with the examples from humanities, where the objects are often bounded and stabilized through social interaction: for literary works, as an example, the boundaries are set through the copyright law and the courts.

Another example that sheds light on social aspect of authenticity is the case of Wikipedia. It is referenced a few times in this article, and it is a common practice now to reference it even in monographs on digital

preservation, e.g. see (Giaretta, 2011). This makes the authors think that Wikipedia is considered a legitimate source of authentic information despite the fact that its articles are edited by a community which is unknown to the user, and can only supply specific evidence of the information trustworthiness.[12] This example confirms the authors' statement made earlier in this study that trustworthiness, or *Warranty,* is not the only aspect of authenticity that the consumer cares about; s/he equally or sometimes more, cares about the *Utility* of the information, and then is ready to partially sacrifice the former in order to gain on the latter.

Authenticity strategies

Even if it is agreed what the basic understanding of authenticity is, it may be questionable how to define an approach, a *strategy* for deciding on what is authentic and what is not. (Rothenberg, 2000) suggests the following authenticity strategies:

- *Originality strategy* that focuses on the originality of the entity; that is, on whether it is unaltered from its original state. Two tactics can be discerned within this strategy:
    - *Intrinsic properties tactics*: to provide criteria for whether each property of the entity is present in its proper, original form. For example, one can demand that the paper and ink of a traditional document be original and devise chemical, radiological, or other tests of these physical properties.
    - *Process tactics*: to focus on the process by which an entity is saved, relying on its provenance or history of custodianship to warrant that the entity has not been modified, replaced, or corrupted and must therefore be original. Intrinsic properties of the entity may be completely ignored using this tactic, since it relies on the authenticity of documentation of the process by which the entity has been preserved as a surrogate for the intrinsic authenticity of the entity.

- *Intrinsic properties strategy* based on the intrinsic properties of the entity but not requiring the properties relation to the originality. This involves identifying certain properties of an information entity that define authenticity, regardless of whether they imply the originality of the entity. For example, one might define an authentic impressionistic painting as one that conforms to the style and methods of Impressionism, regardless of when it was painted or by whom. A less controversial example from (Rothenberg, 2000) might be a jade artifact that is considered "authentic" merely by virtue of being truly composed of jade.

- *Suitability strategy* based on various tactics to specify and test whether the entity fulfils a given range of purposes or uses. This may be logically independent of whether the entity is original.

(Rothenberg, 2000) clearly expressed the inclination to the last sort of authenticity strategy, and pointed out that two other strategies suggested by them do imply some purposes, hence are subcases of a more generic suitability strategy. As an example, if the originality strategy is applied to a certain venerated artifact like the American Declaration of Independence, and such an entity ultimately becomes unsuitable for its normal purpose (such as becomes unreadable), it continues to serve some purpose – in this example, veneration.

(Matthews *et al.*, 2010) and (Matthews *et al.,* 2012) suggest what can be called *behavior strategy* for defining the authenticity that is based on the entity manifestation and behavior. This resembles *intrinsic properties strategy* as it does not imply the originality of the entity; it also resembles *suitability strategy* for if manifestation and behaviour suits a certain pattern, the entity is deemed authentic. The authors, however, consider this a separate strategy focused on the evaluation of the entity's exterior, of its *interface* with the environment. The essence of a behaviour strategy can be illustrated by a metaphor known as Duck Test: *"If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck."*

To discriminate between these strategies, the authors find it useful to talk about a (synthetic) value strategy based on the estimate of value that the Information Entity brings to the Consumer. The value, in turn, is being expressed through the aspects of Utility and Warranty, with parallels with the same concepts from the quality management that were discussed earlier. The value of the entity to the Consumer then becomes a

determinant of the authenticity strategy; the source of value to the Consumer giving relative weight to the authenticity derived from particular aspects, for example between the value derived from originality and the value derived from satisfactory behavioral performance. The value then is just a driver for defining a common approach to authenticity in a particular preservation project; it serves strategic purposes rather that technical ones.

The opposite observation, that the authenticity can be a measure of value, and that the value is maintained by maintaining the authenticity, is also true. The authenticity, however, needs to be thought of in multi-aspect ways according to different strategies outlined above and with finding a good balance that should depend on the nature of a particular preservation project.

The *value strategy* makes most sense, because other strategies can be regarded as just a means of defining value, with some of them inclined more towards the Utility component of it and others more towards the Warranty component. The table below summarizes this observation:

Authenticity strategies	Substantial focus on Utility	Substantial focus on Warranty
Originality strategy		x
Intrinsic properties strategy	x	
Suitability strategy	x	
Behavior strategy	x	
Value strategy	x	x

Table 2: The focus of authenticity strategies on value components.

The value-based approach to authenticity matches the IT Service Management and quality management frameworks which were considered in the first sections of this paper. In addition, the value concept has profound links with social aspects of authenticity that were also touched on earlier, because the value of any digital artifact is to a great extent socially defined.[13] The best way of thinking of it may be that all other mentioned strategies are just *dimensions* of the value strategy. All of them add up to the authenticity "vector" in a sort of a multi-dimensional "authenticity space." For practical purposes of a certain preservation project, it may be useful to focus on one or more of the above introduced strategies, still remembering they are just dimensions of the value strategy.

The variety of authenticity strategies and tactics is represented by the following diagram:

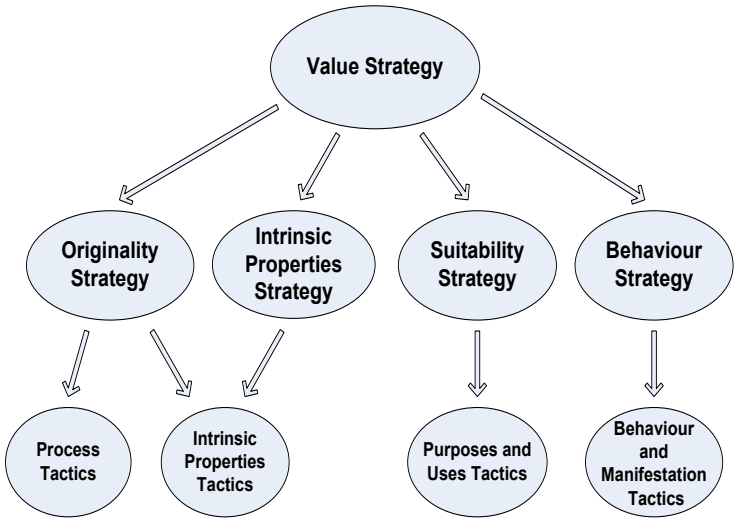

Figure 4: Authenticity strategies and tactics.

Significant Properties as parameters of authenticity model

The literature on authenticity largely intersects with that on Significant Properties of digital information, as the two concepts are naturally related. Good surveys on the matter are (Giaretta *et al.*, 2009) and the appropriate chapter of (Giaretta, 2011), as well as (Knight and Pennock, 2009). These works rightfully mention the disparity of Significant Property definitions. The authors find that, for the purposes of this study, the most valuable definition has been given in (Wilson, 2007):

> *"the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects"*

This definition has its roots in the "performance model" developed at the National Archives of Australia (Heslop *et al.*, 2002) with similar notions discussed at length in (Matthews *et al.* 2010) and (Matthews *et al.* 2012). That model treats the information entity as a result of a mediation of technology and data; the digital entity per se is meaningless until software has performed it for the human; for a concise introduction to these ideas see (Wilson, 2008). The authors appreciate the *focus* of this model: tight conceptual coupling of the data and the software, as well as the necessity of execution/performance of the digital object in order to make it meaningful. The latter aspect is related to the ideas of socially-defined boundaries for information entities that were mentioned earlier, see (Levy, 2000) and (Smith 1996).

On the relation between Significant Properties and authenticity, (Wilson, 2007) cites (Heslop *et al.*, 2002):

> *"neither the source nor the process need be retained in their original state for a future performance to be considered authentic. As long as the essential parts of the performance can be replicated over time, the source and process can be replaced"*

In the OAIS view (see Giaretta *et al.,* 2009), the role of Significant Properties is to support successful transformations (migrations) of Data Objects in other formats (so that transformations preserve Significant Properties and thus authenticity). The term suggested by the OAIS updated version (OAIS, 2012) in place of *Significant Properties* is *Transformational Information Properties*, which emphasizes the supportive if not peripheral role of this concept in OAIS model.

In view of (Wilson, 2007), the role of Significant Properties is more important and is of interest for several American, Australian, and European projects surveyed in the same publication. According to (Wilson, 2007), significant properties of information entities fall into five categories:

- Content, e.g. text, image, slides, etc.
- Context, e.g. who, when, why.
- Appearance, e.g. font and size, color, layout, etc.
- Structure, e.g. embedded files, pagination, headings, etc.
- Behavior, e.g. hypertext links, updating calculations, active links, etc.

This strongly correlates with what was suggested earlier in (Rothenberg and Bikson, 1999) under the name of *authenticity criteria*:

> *"the above intent of these criteria is to ensure that preserved records retain their original behavior, appearance, content, structure, and context, for all relevant intents and purposes"*[14]

The categories seem reasonable and applicable to real cases. If one takes a web page as an example, the Content is represented by text, images, sound and video, as well as values of Ajax objects; Context – by meta-tags and comments; Appearance and Structure – by Cascading Style Sheets; Behavior – by links, JavaScript, Ajax objects, and embedded objects having controls (like audio and video clips).

The authors think that, in addition to those five that originate from the document archiving world, there could be more authenticity categories relevant to specific subject domains that have other, non-documental information entities circulating in them. To decide what authenticity categories are appropriate for a

particular subject domain, the digital preservation expert should consider the *data value drivers*. For facilities science, the data value drivers were analyzed in (Griffin *et al.,* 2012); for other subject domains, there may be other data value drivers such as selection criteria to choose from available authenticity categories. In fact, the authenticity categories can be regarded as parameter "types" of the *authenticity model* of the information entity. The choice of the actual parameters of authenticity model, and metrics for them depends on the nature of the information entity. What is the correspondence between *authenticity strategy* discussed earlier and *authenticity model* introduced now?

Authenticity as discussed earlier has the aspects of Utility and Warranty. The choice of proper dimensions of authenticity strategy helps to *validate* the authenticity of information entities circulating in preservation service. This ensures that "we are doing the right thing" for the Consumer of preservation service, that the right "authenticity vector" or "authenticity function" has been chosen. So the authenticity strategy chosen defines the authenticity strategy in the space of sub-strategies identified, and addresses the Utility aspect of authenticity. As to the proper authenticity model of Information Entities, it defines the authenticity in the space of significant properties, and serves the *verification* purposes when one wants to check that "we are doing the thing right." So the authenticity model addresses the Warranty aspect of authenticity.

Authenticity and Significant Properties from operational research perspective

The realm of authenticity strategies that have been discussed is quite abstract, so any granular consideration or well-defined notation may not be applicable to it. Nevertheless, the suggestion that *value strategy* is related to the rest of authenticity strategies by being dimensions of it may inspire interpretation in the spirit of operational research.[15] Then the implementation and maintenance of preservation service can be considered an *operation* with the purpose of maximizing (or, at least, preserving) the value of data. The "total" authenticity can be thought of as a function of different authenticity sub-strategies taken as "variables;" the purpose of a digital curator then is to maximize this function in time, under certain constraints like costs, available technologies, etc.

In order to model the authenticity and make it measurable, the authors suggested using Significant Properties. Similarly to the authenticity strategy, the authenticity model can be considered from the operational research perspective, too. Just the optimality criterion should be different and look like minimizing the "deviations" or "errors" in the values of model parameters, with some reasonable function that could represent the summary of those "deviations" or "errors" across the preservation period (within the preservation planning horizon) or/and across digital collection.

This operational research perspective is not purely theoretical. The ENSURE project www.ensure-fp7.eu considers optimization of preservation plans for health data and financial information by taking into account the dimensions (aggregated metrics) of the preservation configuration quality, data management costs incurred, and economic models employed. The ENSURE Global Preservation Planning Optimizer component uses genetic algorithms to find Pareto optimal[16] preservation configurations across these three aggregated metrics, and presents them to the Consumer to make a final choice.

More advanced considerations beyond the ENSURE project's scope might lead to seeing the digital preservation operation not only from one Consumer point of view but introducing multiple Consumers as actors. What one might be seeking could be an optimal preservation configuration that satisfies various players in spirit of game theory principles,[17] e.g. there may be some equilibriums or "win-win-win" situations when a few Consumers agree to share a certain preservation configuration, or a set of them. An even more complex setting could be considered if a preservation solution (digital archive) manager is introduced as an additional player having her own interests, like making profit out of the digital preservation service.

Data collections in long term digital preservation environments

Once generalized, the authenticity can be considered closely correlated to data value, and be used in the preservation policy context. The data authenticity and data value then become important drivers, if not determinants, of a sound preservation policy. As the authors are concerned with the long term, the models of

data authenticity and data value are subject to change; hence the preservation policy should evolve accordingly. The authors discuss the use of this framework in a scientific collection management context.

An organization does not collect individual items/objects but has the concept of a collection or content set which is linked to the purpose and mission of the organization. All collections of information, whether they comprise physical or digital objects, need to be acquired or created, collected, described, kept & managed, used and disposed of (or transferred). Different disciplines and purposes will put different weights on these stages and will have different standards and constraints to apply, and factors which are encapsulated in a preservation policy. This policy and its underlying assumptions may be implicit or explicit. These collections are not kept in a vacuum; they are there to satisfy the needs of a clearly identified Designated Community. This approach is a Utility based one, putting the needs of the Designated Community at the centre of collection and service developments. The role of the organisation affects when the value and success of that acquisition can be measured.

The overall purpose of collection or data management processes is to maximise the value of the collection to the Designated Communities or subsets of the Community within the wider whole. It can be done through the variety of digital curation practices; as an example, minting persistent identifiers for data can be seen as a means of raising its value, see (Wilson, 2012).

For a library collection the material goes through a series of value phases. Initially when a specific item is acquired it has high value to the service and its users as the content is new and up to date; over a period of time this content will be superseded and the value of this item decays; following a further period of time the value can increase again from an historical context and may in fact have become rare due to others disposing of their copies; however, for a working library it is likely that this historic value may not be of value to their Designated Community but may be to others in the wider environment.

As a concrete example, the authors are working with the STFC ISIS Neutron Facility[18] to support the preservation of their collection of data arising from the use of the facility in scientific experiments on material samples (for example crystals, chemical compounds, biological samples, or engineering components). As ISIS is a working scientific facility producing data and supporting analysis, there is not the same concept of "collection management" as the resulting data is not the end-point but a stage in the wider scientific lifecycle. Thus the aspects of collection management are extracted from the policy documents concerning the management of the facility. Key in this is the data management policy.[19]

The ISIS scientific data comes from a different perspective from an institution which collects content, the "acquisition" decisions are made at the proposal stage of the business process – is the experiment of sufficient scientific value to be given beam time on the ISIS instrument? The raw data and some automated metadata are then automatically collected and the immediate retention decision is "forever." It is a policy decision to retain data rather than to take the effort to identify and remove poor quality data.

Classification of the collection is a key tool in identifying value within a library collection, by distinguishing between materials, parts of the whole collection may be identified as important to the user community, rare or expensive to replace and thus increase in value. The ISIS data is not classified in this way – it is currently functionally classified by collection date; however, there are other possibilities which might help tease out different levels of value within the whole collection. For example, for ISIS data one can discern the data format, and importantly, also the context in which the data has been collected. This includes the instrument used and experiment undertaken; the material sample studied; the year it was undertaken in and going forward the type of analysis to be undertaken. By classifying data by the context, the authors increase the provenance of the data (and thus its authenticity) and thus the value for further processing, re-examination, and reuse.

Taking a different perspective and thinking about the concept of organising the data to support preservation actions; the business process could be related to the instrument and experiment so that changes can be linked to the software modules, instruments and experiments also provide value as the type of analysis and experiment undertaken. There might be, for a domain specialist, some hierarchy of scientific worth related to instruments – so that newer or more precise instrumentation might make data from some experiments of more long-term value than others. The material sample used is also an important factor in establishing value

– it may have rarity value due to the complexity of producing it or in the risks associated with its properties and these differences can also affect the notional value of the data.

Rarity in general collection management can come in two forms – rare from the start as not many were/can be produced so that rarity value is established from the start; or rare because it has survived where lots of similar items haven't as the item didn't have great value associated with it to start with – manuscripts are a good example of the former and toys and comics of the latter.

Rarity from the start can be built into the acquisition process and can be exhibited by different storage/description arrangements from the initial stages. For the traditional library domain, the judgement on rarity can be based on the librarian's expert knowledge of the field being acquired; for digital libraries holding research data, this could be identified at the stage when the research proposal is reviewed, and then flagged up in the associated metadata. It is much harder to assess the second kind without putting in place a new process to look for this, via assessment of the alternates available for example and the additional representation information, for example, might not be available at that later stage – just like an old toy may no longer have its box!

The data collections aspects considered are in fact *drivers* for data value in facilities science domain. For other domains: data archives and libraries, or business and industry, the value drivers may be different. The Consumers (Designated Communities) may differ, too, with an indication of possible variety of them, again for the case of facilities science, outlined in (Wilson, 2012).[20] What is going to remain a permanent theme for all digital preservation domains however is the notion of value and its relation to authenticity; in the absence of sound economic models, or where they are not easily applicable owing to the nature of the digital collections, the wider understanding of authenticity and its relation to value should serve the design of a sound preservation policy.

Preservation policies for collections and organizations

To be able to manage collections effectively and to design and select the appropriate long term digital preservation solution there must be the appropriate policy framework in place; this framework should include policies concerned with the preservation of the objects within the collection.

In view of the value-based approach to authenticity, the digital preservation policies should then be considered an important input to the data authenticity modeling and should, in turn, incorporate the notion of data value, and the need of optimizing it through time. Speaking of preservation policies, there are at least three layers considered in the SCAPE project www.scape-project.eu:

- Preservation Policy or Guidance Policy or Preservation Strategy. This is typically a high level document which sets out the general approach and ethos for the preservation. It is written in natural language with a target audience of other humans; for example, there may be a general statement about using well defined formats for digital objects.
- Preservation Procedure Policy. This document is more detailed than the Guidance policy, but is still pitched at a general level. Taking the file format example, the document will go into more details about well-defined means but will not put concrete file types in. This is also a natural language document intended to be read by other humans.
- Actionable Preservation Policy, or control policies. Policy at this level is concerned with specific collections and will be created in both a human and machine actionable formats. This machine policy can be used in preservation planning and watch tools to ensure that items of importance are checked for and taken into account during preservation planning. At this level the significant properties of the digital object and associated collection are of paramount importance, as this level of policy sets the relative priorities of different properties of the object.

For authenticity, the top-level Preservation Policy sets out the general aims. Preservation Procedural Policy should prioritize sub-strategies that are outlined in this paper, and their contribution to the major authenticity strategy "vector." The scope of the Actionable Preservation Policy is the selection of data authenticity

categories (significant properties) as well as of particular variables and metrics for them which serve as monitoring parameters of authenticity and trigger necessary actions in order to prevent the data value loss through time.

The data authenticity and data value then become important drivers, if not determinants, of a sound preservation policy. As the authors are concerned with the long term, the models of data authenticity and data value are subject to change; hence the preservation policy, and digital collections guided by the policy should evolve accordingly.

Conclusion

The authors placed the *preservation solution* in the context of *preservation service*, and the latter in even a wider context of a generic IT service. This allowed the authors to apply well-known IT Service Management and quality management frameworks and identify supplying *value* with its aspects of Utility and Warranty as a generic objective of preservation service. The authors then considered the Access and Authenticity aspects of handling the information entities, and identified these aspects as being yet another break-down of IT Service that is specific to the domain of digital preservation. The authors focused on the authenticity aspect, revisited the existing conceptual analysis in the field, and agreed on *value authenticity strategy* as the most generic one that can incorporate other authenticity strategies. The top-down conceptual analysis from the IT Service Management perspective and the bottom-up considerations of authenticity in a large corpus of earlier research then met, and led the authors to the conclusion that authenticity and value notions are tightly coupled. The authors considered the role of collections in digital preservation, and drivers for the value of them. The authors then looked at the preservation policies and various layers of them, and suggested the data value and data authenticity to be important factors in the design, the actual implementation, and the evolution of collections driven by the policies.

The authors consider some themes that are only touched on in this paper; interesting areas for further conceptual analysis, as well as for the actual design of preservation solutions and services. Social boundaries of digital objects may be one of these themes: how one defines the boundaries, how one models them (probably with some semantic and Linked Data techniques), and how one makes the socially defined digital objects operable by human and machine agents – this may constitute a subject of a separate thorough study or a project. Sensible modeling of preservation policies through all three levels that are mentioned: Guidance Policy, Procedural Policy, and Actionable Policy, the validation of true correspondence among these levels, as well as the design of protocols for the policies execution and verification – can be another fruitful area of research. Modeling the priorities of Designated Communities and other Consumers of preservation services in spirit of operational research and game theory, then applying these models for managing large digital collections can be the third direction of research where this study may contribute.

References

Authenticity (2000), *Authenticity in a Digital Environment.* Council on Library and Information Resources. Washington, D.C.

Beagrie, N., Chruszcz, J., and Lavoie, B. (2008), *Keeping Research Data Safe: a cost model and guidance for UK universities*, Final Report April 2008, available from http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf (accessed July 10, 2013).

Beagrie, N., Lavoie, B., and Woollard, M. (2010), *Keeping Research Data Safe 2*, Final Report April 2010, available from http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf (accessed July 10, 2013).

Conway, E., Giaretta, D., Lambert, S., Matthews, B. (2011), *Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context.* The International Journal of Digital Curation. Vol 6, No 2.

Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G., Guercio, M. (2009), *Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage.* TaPP'09. The First Workshop on the Theory and Practice of Provenance. San Francisco, February 23, 2009. http://static.usenix.org/event/tapp09/tech/full_papers/factor/factor.pdf (accessed July 10, 2013).

Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercio, M., Michetti, G., Sawyer, D. (2009), *Significant Properties, Authenticity, Provenance, Representation Information and OAIS (2009).* Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects. http://escholarship.org/uc/item/0wf3j9cw (accessed July 10, 2013).

Giaretta, D. (2011), *Advanced Digital Preservation.* Springer.

Griffin, T., Matthews, B., Mills, A., Nagella, S., Shaon, A., Wilson, M.D., Yang, E. (2012), *Moving from a scientific data collection system to an open data repository.* Proc. The 7th International Conference on Open Repositories (OR2012), Edinburgh, UK, 09-12 Jul 2012. http://epubs.stfc.ac.uk/work-details?w=62492 (accessed July 10, 2013).

Guercio, M. (2008), *Authenticity and OAIS. The CASPAR model and the INTERPares principles & outputs.* Delos Summer School. June 2008. http://www.slideshare.net/DigitalPreservationEurope/authenticity-and-oaisthe-caspar-model-and-the-interpares-principles-outputs-presentation (accessed July 10, 2013).

Guercio, M., Michetti, G., Meghini, C. (2009), *Modeling Authenticity.* DPE preservation training materials. March 2009. http://www.digitalpreservationeurope.eu/preservation-training-materials/files/authenticity.pdf (accessed July 10, 2013).

InterPARES (2002), *InterPARES Authenticity Task Force report.* http://www.interpares.org/book/interpares_book_k_app02.pdf (accessed July 10, 2013).

Heslop, H., Davis, S., Wilson, A. (2002), *An Approach to the Preservation of Digital Records*, National Archives of Australia. http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf (accessed July 10, 2013).

ITIL Strategy (2007), *ITIL. Service Strategy*. OGC. London: TSO.

ITIL V3 (2007). *ITIL V3. A Pocket Guide.* Van Haren Publishing.

Knight, G. (2009), *Data Witout Meaning: Establishing the Significant Properties of Digital Research.* International Journal of Digital Curation, issue 1, Volume 4, 2009.

Kiefer, S., Wilson, M.D. (2012), *Ensuring profitability of commercial long term digital preservation.* ERCIM News 91 (Oct.), pp. 19-20. http://epubs.stfc.ac.uk/work-details?w=63731 (accessed July 10, 2013).

Knight, G. and Pennock, M. (2009), *Data Without Meaning: Establishing the Significant Properties of Digital Research.* International Journal of Digital Curation, issue 1, Volume 4.

Levy, D. (2000), *Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment.* In Authenticity (2000).

Lynch, C. (2000), *Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.* In Authenticity (2000).

Matthews, B., Shaon, A., Bicarregui, J., Jones, C. (2010), *A Framework for Software Preservation.* The International Journal of Digital Curation, 2010, Vol. 5, No. 1, pp. 91-105.

Matthews, B., Shaon, A., Conway. E. (2012), *How do I know that I have Preserved Software?* In *The Preservation of Complex Objects. Volume 1 Visualisations and Simulations*, eds. Janet Delve, David Anderson, Milena Dobreva, Drew Baker, Clive Billenness, Leo Konstantelos (The University of Portsmouth), pp. 36-53.

OAIS (2012) *Reference Model for an Open Archival Information System (OAIS)*. Magenta Book. Issue 2. June 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf (accessed July 10, 2013).

Pugh, K. (2006), *Interface-Oriented Design.* The Pragmatic Programmers LLC.

Rothenberg, J. and Bikson, T. (1999), *Carrying Authentic, Understandable and Usable Digital Records Through Time. Report To the Dutch National Archives And Ministry of the Interior.* RAND/RE-99-016. http://www.rand.org/pubs/rand_europe/RE99-016.html (accessed July 10, 2013).

Rothenberg, J. (2000), *Preserving Authentic Digital Information.* In Authenticity (2000).

Smith, B. C. (1996), *On the Origin of Objects*. MIT Press, Boston.

Wilson, A. (2007), *Significant Properties report.* InSPECT Work Package 2.2, Version 2. April 2007. http://www.significantproperties.org.uk/wp22_significant_properties.pdf (accessed July 10, 2013).

Wilson, A. (2008), *Significant Properties of Digital Objects.* Presentation on the JISC, the British Library and the Digital Preservation Coalition (DPC) joint workshop at the British Library Conference Centre. April 2008. http://www.dpconline.org/component/docman/doc_download/142-presentation-wilson (accessed July 10, 2013).

Wilson, M.D. (2012), Meeting a scientific facility provider's duty to maximise the value of data. In *DataCite Summer Meeting, Digital Research Data in Practice (DataCite2012), Copenhagen, Denmark. http://epubs.stfc.ac.uk/work-details?w=62852* (accessed July 10, 2013).

---

[1] SCAPE: SCAlable Preservation Environments. http://www.scape-project.eu (accessed July 11, 2013).

[2] ENSURE: Enabling kNowledge Sustainability Usability and Recovery for Economic value. http://www.ensure-fp7.eu (accessed July 11, 2013).

[3] SCIDIP-ES: SCIence Data Infastructure for Preservation – Earth Science. http://www.scidip-es.eu/ (accessed July 11, 2013).

[4] *Consumer* is a legitimate OAIS term (OAIS, 2012) and a concept that is a generalization of *Designated Community*.

[5] There is a fairly often used term *preservation objective* which is very granular and concrete, see (Conway *et al.*, 2011). In this paper, the *preservation service objective* designates a different high-level concept applied to the service as a whole.

[6] ITIL official Website: http://www.itil-officialsite.com/ (accessed July 11, 2013). ITILv3 underpins ISO/IEC 20000 (previously BS15000), the International Service Management Standard for IT service management, although differences between the two frameworks do exist.

[7] ITSMWatch on Utilility and Warranty. See http://www.itsmwatch.com/itil/article.php/3863596/How-to-Measure-ITIL-Service-Utility-and-Warranty.htm (accessed July 11, 2013).

[8] ISO 9000 series of standards. http://en.wikipedia.org/wiki/ISO_9000 (accessed July 10, 2013).

[9] Business Dictionary definition of "fitness for use". http://www.businessdictionary.com/definition/fitness-for-use.html (accessed July 11, 2013).

[10] In OAIS's view with which we agree it would be more correct to say that what circulates in the information lifecycle is *information packages.* For the moment, it is just more suitable to omit the packaging layer and discuss *information entities* (a generalization for *information objects*) straight away as this is what Consumer cares about, and what Access and Authenticity aspects relate to.

[11] CASPAR – Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. http://www.casparpreserves.eu (accessed July 10, 2013).

[12] Wikipedia does have its own guidelines on the quality of content, also its own social mechanisms in support of its mission. We appreciate all this; we only want to emphasize the *specifics* of these mechanisms that are, simultaneously, more open and more anonymous than those in case of a traditional encyclopaedia like Britannica.

[13] The idea of socially constructed digital objects (akin to traditional artifacts of cultural importance) should not be considered something new. IT professionals commonly use socially defined constructs, often without realizing it. Data types, programming languages, Application Programming Interfaces, or software development frameworks are just specifically formulated social contracts. Not humans only but machines, too, can be a partner in such a contract.

[14] As cited by (Giaretta *et al.*, 2009).

[15] Operational research, or operations research. https://en.wikipedia.org/wiki/Operations_research (accessed July 25, 2013).

[16] Pareto efficiency. http://en.wikipedia.org/wiki/Pareto_efficiency (accessed July 10, 2013).

[17] Game theory. http://en.wikipedia.org/wiki/Game_theory (accessed July 25, 2013).

[18] www.stfc.ac.uk/ISIS

[19] ISIS Data Management Policy. http://www.isis.stfc.ac.uk/user-office/data-policy11204.html (accessed July 10, 2013).

[20] And with each consumer potentially having her own understanding of value – which brings again the game theory perspective.