# A Framework for Applying the Concept of Signi cant Properties to Datasets

**Simone Sacchi, Karen Wickett, Allen Renear, and David Dubin**
{sacchi1, wickett2, renear, ddubin}@illinois.edu
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211 USA

## ABSTRACT

The concept of *signi cant properties*, properties that must be identified and preserved in any successful digital object preservation, is now common in data curation. Although this notion has clearly demonstrated its usefulness in cultural heritage domains its application to the preservation of scientific datasets is not as well developed. One obstacle to this application is that the familiar preservation models are not sufficiently explicit to identify the relevant entities, properties, and relationships involved in dataset preservation. We present a logic-based formal framework of dataset concepts that provides the levels of abstraction necessary to identify and correctly assign significant properties to their appropriate entities. A unique feature of this model is that it recognizes that a typed symbol structure is a unique requirement for datasets, but not for other information objects.

## KEYWORDS

Dataset, Significant Properties, Digital Preservation, Data Curation, Preservation Models, Ontology, Conceptual Foundations.

## INTRODUCTION

The concept of significant properties, properties that, in some given context, must be identified and preserved in any successful digital object preservation, is now common in data curation and preservation (Hedstrom & Lee, 2002; Hockx-Yu & Knight, 2008; Brown, 2008; Knight, Grace, & Montague, 2008; Matthews, McIlwrath, Giaretta, & Conway, 2008; McDonough, 2011). This notion has demonstrated its practical usefulness in cultural heritage domains, and exploration of applications to scientific datasets is now underway. However as Giarretta et al. (2009) have shown, the definitions of significant property vary widely. Moreover they are in almost every instance too informal and colloquial to support a systematic understanding.

*ASIST 2011*, October 9–13, 2011, New Orleans, LA, USA.



Part of the problem is that the common discourse of digital preservation appears to be fundamentally metaphorical and so can be deeply misleading as to the precise nature of digital preservation. We routinely speak of *preserving something* — and yet what is the thing that is preserved? The apparent ontological commitments of digital preservation discourse are still based on a metaphor of physical preservation, where it is plausibly the continued existence of a physical object, or the maintenance of the features of a physical object that is the goal of preservation actions.

But this is not what is going on in digital preservation, or at least not primarily. There is some accommodation in our common preservation discourse of the subtleties of digital preservation. It is, for instance, a commonplace that maintaining the physical existence of a CD-ROM, or maintaining certain aspects of its physical condition, is only a part of the preservation problem for the information on that CD-ROM. And one hears slogans such as: *preserve the information, not the bits*. But the problems caused by the dominant metaphor are deep and entrenched and cannot be resolved without a much more formal and systematic theory of preservation than we have now.

Consider the need to preserve a recorded bit sequence of interest, and re-present that same bit sequence in new media with some warrant of authenticity and integrity. Even here, at this relatively low "physical" level we already encounter conflict with the metaphor of preserving the relevant characteristics of a physical object. The particular bit sequence that is "preserved" in migration to new media is not itself a physical object. It is a single abstraction that can be repeatably and multiply instantiated in various physical things. Consequently there is no (literal) sense in speaking of *preserving* a bit sequence as if it were an object subject to corruption. In particular making physical copies of a physical object does not in any way *preserve* that bit sequence; rather those copies are entirely new instantiations of the same bit sequence. It is at least an idiom and at worst a misleading "category mistake" to speak of preserving bit sequences — their persistence is ensured by their fundamental nature as repeatable abstractions (rather than concrete physical objects) and so preservation is neither required, nor possible.

Much current focus in digital preservation is at levels above

both the physical objects and the lower level abstractions (such a bit sequences) instantiated in those objects. But, for the same reasons, the rhetoric of "preserving meaning" or "preserving information" is strictly speaking, false, and if intended as a metaphor, profoundly misleading. Those higher-level things (meaning, information, properties) also, as repeatable abstractions, require no preservation efforts on our part to ensure their continued existence. With respect to significant properties, Dappert and Farquhar have in fact already remarked, "It is . . . not sensible to speak about preserving a digital property" (2009).

This may seem an uncharitable reading of contemporary preservation discourse, a straw man even. It might be argued that we are being willfully literal-minded and that having recognized these certain expressions as idioms we should move on and focus on what is intended by them and not their surface grammar. Obviously what is meant by "preserving significant properties" is not preserving the abstract property itself, but ensuring that the object of interest, undergoing preservation, continues to have that property. Unfortunately this does not solve the problem. In many cases (of migration) the object that had the property does not even exist after a successful migration. Not that it is even clear what sort of "object" this is.

It is our contention though that the logic of preservation, and especially digital preservation, is so challenging that idiom and metaphor systematically impede progress and only careful literal expression will help us advance. ER and UML diagrams are a step in the right direction, but a better tool for making preservation models explicit enough to help us with the hard problems of digital preservation theory, is axiomatic expression in formal logic.

These issues are not new (Thibodeau, 2002), nor is the call for a formal theory of preservation (Cheney, Lagoze, & Botticelli, 2001; Flouris & Meghini, 2007) that would proved a more perspicuous language of preservation. But a shared conceptual framework that is relatively free from misleading idioms has not yet been realized.

In what follows we adapt one effort to develop a formal ontology of dataset concepts (Dubin, 2010) to the problem of understanding how the notion of significant properties can be applied to datasets.

## GENERAL APPROACH

At the most general level we, like Mois et al. (Mois, Klas, & Hemmje, 2009) understand digital preservation as being, or perhaps better, as ensuring, "communication with the future". Though of course the activities described as preservation typically reflect a concern that some special attention is needed if the intended communication is to succeed.

Some communication, digital or otherwise, has a particular restricted purpose: providing information. The preservation of datasets is communication of this sort. Informative communication is accomplished by manipulating physical objects and events so that they instantiate signs, which then, in a context with the appropriate social conventions and practices, express propositions, and result in someone recognizing, with some level of warrant, that the intended propositional content *is* the intended propositional content. The technology may vary in nature and complexity, but the basics of the process are the same. The conceptual model that we develop below reflects this account of how communication takes place, whatever the mechanism or technology.

Part of our strategy for ensuring that we are identifying objects correctly, and avoiding conflation or misidentification, involves systematically distinguishing: [1] idiomatic from literal predication, [2] individual concrete things from repeatable abstractions, and [3] types of things from roles that types of things enter into in certain circumstances.

We illustrate the first two distinctions with an example inspired by FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998). Consider a single copy of a printed book. There are ink marks on the page; there are graphemes that are rendered by those marks; there are sentences that are expressed by those graphemes, and there is a story that is told by those sentences.

**Idiomatic vs Literal Predication:** There are four things here, the story (which might be told in many different languages), the sentences (which could be written or spoken.), the graphemes (which admit some variation in size and shape), and the individual physical book with its ink marks. Each thing has its own unique and disjoint properties. We might say that the physical book is soiled, in English, and mostly false. But this is not literally true; what is *literally* true, and what we mean, is that the book is soiled, the sentences are in English, and the story those sentence tell is mostly false.

**Abstract vs. Concrete:** the physical book is an individual concrete thing. It exists for a period of time and in a connected series of locations, enters into various causal relationships, and undergoes change (losing parts, fading in color, etc). The sentence and the story are not individual concrete things. They do not, at least not in the same way, exist in space and time or enter into causal relationships. The story and the sentences that tell the story are both abstract things. Abstract things are *repeatable*; they have *instances*: the story can be told over and over, the sentence can be uttered again and again. Individual concrete things cannot be repeated and do not have instances (they *are* instances)[1]. The commonsense of "abstract," already in evidence in preservation standards documents, is enough for our purposes.

**Types vs Roles:** Here we use the standard example. Student, senator, author, etc. are all roles that a person enter into, and can leave. A person can become a student, and then later cease to be a student. But person is not like that: a person cannot cease to be a person (although, of course,

[1]We do not intend a philosophical position here; we do not claim that abstract objects are real in any deep metaphysical sense, or that they cannot be reduced to some other sort of objects (concrete things, practices, beliefs, etc.).

| | | | | |
|---|---|---|---|---|
| 77 | 167 | 84 | 242 | 570 |
| 11 | 23 | 2 | 16 | 52 |
| 16 | 33 | 3 | 3 | 55 |

**Table 1. Example of Symbol Structure**

a person can cease to be). More generally, a student might not have been a student, but a person cannot have been anything other than a person. We say that person is a type and student is a role. This powerful distinction, presented in simplified form here, has been shown to provide deep insights into how to shape and simplify conceptual models and ontologies (Guarino & Welty, 2000).

We will see that these three critical distinctions can disentangle some of the intricacies and puzzles involved in understanding the logical nature of digital preservation.

## THE CONCEPT OF DATASET

We understand a dataset as a symbol structure that, in the context of an assertion event, expresses data content and supports certain kinds of operations, assigning values to some observable phenomenon on the basis of an observational or computational process. While there is ontological disagreement over the entities, properties, and events involved in the observation process (Cox, 2006; Madin et al., 2007; Kuhn, 2009), there is general agreement about the nature of recorded data as the symbol-mediated assignment of values (quantitative, qualitative, or categorical) to observed phenomena.

These concepts are based on a formal account of data that came out of recent digital preservation research (Dubin, 2010; Sandore & Unsworth, 2010; Dubin, Futrelle, Plutchak, & Eke, 2009). The present theory supports an intuitive notion of data as something recorded as a result of an observational process and recognizes that the same data can be serialized in different file formats. A dataset, on this account, is a symbol structure that expresses propositional content (data content).

As an example of dataset consider Table 1, a two dimensional array (a symbol structure) of numerals. This table expresses the number the students enrolled at a graduate school for the academic year 2009-2010, each row respectively representing the number for the Masters Program, the Certificate of Advanced Studies (CAS) Program, and the Ph.D Program. The first two columns express the number of full-time students, respectively male and female; the third and fourth columns express the part-time students, respectively male and female.

The symbol structure by itself has no semantic meaning and is therefore not, by itself, a dataset. It's the connection with specific data content for the purpose of making a statement about enrollment that makes this symbol structure a dataset.

When a symbol structure expresses data content in some context, we can say that it is in a data content bearing *role* (Guarino & Welty, 2000). This definition of a dataset is similar to accounts of digital documents as abstract kinds that can be serialized in various forms, and, more specifically, parallels Renear and Wickett's (2009) treatment of documents as "strings in a role". But should we say that if a symbol structure expresses data content it is therefore a dataset?

Perhaps not. While bearing data content is a necessary condition for a symbol structure to be a dataset, it does not give us the whole picture. There is an additional requirement for a symbol structure to qualify as a dataset: they should support the ability to perform certain kinds of operations with the data.

A symbol structure that is a dataset has a particular structure and type (or several types) which govern the operations that could be performed over the data. This is comparable to the notion of data type in programming languages however we warn against taking the analogy too far. The definition of a precise syntax for the symbol structure has been identified as a core feature in the definitions of dataset in scientific literature and technical documentation (Renear, Sacchi, & Wickett, 2010). This is critical if the identity conditions we establish for datasets are to signify not only the identity of data content, but also identity with respect to the supported operations which we believe to be part of the common concept of dataset, but not accommodated in accounts such as Dubin's (2010).

For instance, there are many ways one might express the proposition that seventy seven male students enrolled in our Masters program in that academic year. Notation like the strings of Arabic numerals in this table influence both how readers interpret the information, and what kinds of operations on the data are easy or hard to perform. For example, if the cardinal number were expressed using a row of black and white wooden cubes (`bwwbbwb`) then multiplication by two could be accomplished with a simple shift of position, just as multiplication in base ten is accomplished by appending a zero to the right side of the string. The recording medium (ink on paper, wooden tokens, etc.) also places constraints, but we simply assume some form of digital representation with positive and reliable methods for reading and writing discrete symbols (Haugeland, 1981).

Scientific data is typically recorded with the aim of manipulation by electronic computers, with one or more interpretive layers between the binary string and the notation in which a scientist asserts a proposition. As with the paper/cube example, these digital formats constrain the kinds of operations that are convenient to perform. The numbers in the table might be digitally expressed in two's complement binary format, as strings of ASCII numerals, as a vector graphic of numeral outlines, or as a raster image, with different implications in each case for the complexity of arithmetic or other transformations of the data. Operations of software tools (programming languages, office applications, etc.) are usually governed by some classification of data types, such as integer, character string, array, function, or geometric shape. We abstract away from the details of type classification differences between tools, and regard dataset type as an assignment from any well-defined type classification, with rules

specifying permissible operations. This example provides an idea of how both the operations and content of a dataset contribute to the scientific identity of datasets.

Being in a dataset role is a contingent relational property of a symbol structure, and is assigned by some scientific community. The same symbol structure could express different data content or be typed differently in different scientific communities.

We introduced the notion of Intended Community to identify the community for which a particular symbol structure expresses specific Data Content and should support specific operational capabilities. Intended Community is similar to the concept of Designated Community in OAIS ("Reference Model for an Open Archival Information System (OAIS)", 2002). Dataset properties can only be identified with respect to the Intended Community as they are bound to the community interpretation and use of a symbol structure: they depend on the conventions and expectations of the Intended Community. Both Data Content and Dataset Operation are components of a specific community-related dataset identity.

Digital datasets are encoded in a digital format and are serialized as digital objects. A dataset, as an entity, could not be identified with a particular digital object, not even for the purpose of preservation.

Several of the properties ascribed to datasets are not properties of digital objects: for example being an array of natural numbers is not a property of a digital object itself but of a higher level symbol structure (the one having the role of being a Dataset). There are properties at several levels of abstraction that are important for the interpretation and use of a dataset, so we need to discriminate between these levels in order to correctly assign properties as we prepare for preservation.

## A CONCEPTUAL MODEL

Further clarification of the notions introduced so far is needed in order first to show how the concepts in existing preservation models — OAIS ("Reference Model for an Open Archival Information System (OAIS)", 2002) and PLANETS (Farquhar & Hockx-Yu, 2008) in particular — relate to ours, and then to show how the concept of significant properties can be given a solid foundation. We begin first with natural language definitions, and then re-express each in first order logic.

**Dataset Content:** the propositional content identified by the Intended Community for a Dataset. The information needed to connect a symbol structure to a particular Dataset Content are expressed by an Expression Model.

**Expression Model:** a set of rules that maps propositional content to a particular symbol structure.

**Dataset Operations:** the set of operations identified by the Intended Community as appropriate for a Dataset. The set of operations are determined by the Dataset Type assigned by the Intended Community to a particular symbol structure.

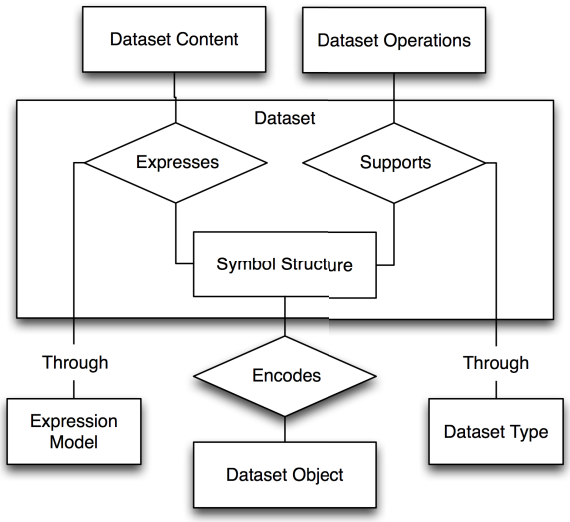


**Figure 1. Conceptual model diagram.**

**Dataset Type:** a classification of symbol structures that determines the Dataset Operations. The Dataset Type is also assigned by the Intended Community.

**Dataset:** the primary symbol structure for a systematic assertion, i.e., an assertion justified by observation or computation (Dubin, Wickett, & Sacchi, 2011). For the Intended Community it [1] expresses Dataset Content, and [2] supports operations appropriate to its Dataset Type. These are all necessary conditions for a symbol structure to be a Dataset.

**Dataset Object:** an encoding of a dataset as a digital object in conformance to specific representation schemes and file formats.

Axioms relating these entities:

[1] Dataset Content is propositional in nature. Propositions can be arbitrarily complex, and so the content of a dataset can be understood as a single proposition:

$$\forall x[\, DatasetContent(x) \;\Rightarrow\; Proposition(x)\,] \qquad (1)$$

[2] A Dataset is a symbol structure that, for an Intended Community, expresses Dataset Content and supports particular Dataset Operations:

$$\begin{aligned} &\forall x \forall w \{ Dataset(x,w) =_{df} \qquad (2)\\ &\quad SymbolStructure(x) \wedge IntendedCommunity(w) \wedge \\ &\exists y \exists z [\, DatasetOperations(z) \wedge DatasetContent(y) \wedge \\ &\quad Expresses(x,y,w) \wedge Supports(x,z,w)]\} \end{aligned}$$

[3] A Dataset Object is a digital object that encodes a Dataset y for an Intended Community:

$$\forall x \forall w \{ DatasetObject(x,w) =_{df} \quad (3)$$
$$DigitalObject(x) \wedge IntendedCommunity(w) \wedge$$
$$\exists y [\, Dataset(y,w) \wedge Encodes(x,y,w) \wedge ] \}$$

This simple conceptual model provides abstraction layers that discriminate properties governing the scientific identity of a Dataset for an Intended Community. It connects observation models (Cox, 2006; Madin et al., 2007; Kuhn, 2009) with preservation models such as OAIS and PLANETS and provides a framework to precisely identify significant properties and assign them to the entity to which they really belong. This model, an adaptation of Dubin's (2010) has obvious parallels with FRBR (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998), with Data Content corresponding to Work, Dataset to Expression, and Data Object to Manifestation. For an attempt to model datasets directly with FRBR see Hourclé (2008).

## PRESERVATION TARGET VS. PRESERVATION OBJECT

For a correct assignment of significant properties we need to distinguish between preservation target and preservation object. The preservation target is the entity we want to ensure access to. The ultimate goal of preservation actions is to maintain this access. The preservation object, on the other hand, is the entity against which the preservation actions are performed. This distinction helps resolve ambiguities in current discourse around preservation.

In our account, the preservation target is the Dataset, a symbol structure in a role. It is not enough to preserve the symbol structure itself — as we noted in the introduction it is hard to understand what it would mean for a symbol structure itself to be preserved or not preserved. It is the symbol structure being in a particular role — expressing Dataset Content and supporting Dataset Operations — that must in some sense be preserved.

However, roles in the Guarino and Welty sense (2000) cannot themselves be the preservation target. They need a proper kind that plays that role. Our kind is, of course, the symbol structure. The symbol structure is encoded in a carrier, usually a set of bits to be preserved in an actual information system. The Dataset Object performs this function in our model and we therefore consider it our preservation object. On this account, provenance information (Moreau et al., 2008) for the preservation object has to be maintained to ensure a correct encoding of the preservation target over the chain of preservation actions. Preservation actions often consist of migrating the preservation target from one encoding to another. The preservation object in fact, may not survive a preservation action — or it may survive, but becomes irrelevant — because the output can be a new digital object. Although preservation is directed at the preservation object, it is not an attempt to preserve the preservation object.

We believe this distinction to be important for the identification of significant properties. In many cases of "preserving significant properties" what is being preserved is not the state of some object having that property, but rather the state of some object being related to another object that has that property. So if F is a significant property of the Dataset which is the Target of Preservation, then the significant property F is preserved, not by ensuring that the Dataset continues have F, but rather by modifying (or providing a new) an Object of Preservation that, for the intended community, encodes the Dataset that is the Target of Preservation. So if F is a significant property of a dataset, F is preserved by providing a preservation object that, for the intended community, correctly encodes that dataset.

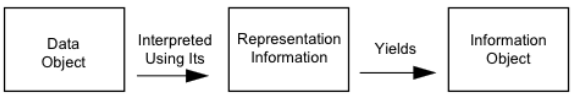


**Figure 2. Obtaining Information from Data in OAIS.**

## COMPARISON WITH OAIS AND PLANETS

Can our more ontologically precise conceptual model be coordinated with influential but less explicit preservation models such as OAIS and PLANETS?

### The OAIS Reference Model

Section "2.2.1 Information Definition" of the OAIS Reference Model ("Reference Model for an Open Archival Information System (OAIS)", 2002) defines "Information" as:

> "any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data. For example, the information in a hardcopy book is typically expressed by the observable characters (the data) which, when they are combined with a knowledge of the language used (the Knowledge Base), are converted to more meaningful information."

with this additional example:

> "the information stored within a CD-ROM file is expressed by the bits (the data) it contains which, when ...combined with the Representation Information for those bits, are converted to more meaningful information as long as the Representation Information is understandable using the recipient's Knowledge Base."

It is very difficult to interpret this definition literally. For OAIS "Information" seems to be propositional in nature. In addition OAIS recognizes symbol structures such as observable "characters" and "bits."

The particular kind of Information Object that is the target of preservation in OAIS is the Content Information, defined as follows:

> "The set of information that is the original target of preservation. It is an Information Object comprised of its Content Data Object and its Representation Infor-

mation. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc."

This definition seems to assume that an Information Object is a tuple — Data Object and Representation Information — or is composed of Data Object and Representation Information as implied by the open diamonds that connect Data Object and Representation Information in the OAIS Information Model diagram in Figure 3. The diagram in Figure 2 and the one in Figure 3 seem to be inconsistent. This problem is reflected in the following example taken again from the OAIS Reference Model:

> "assume the bits represent an ASCII table of numbers giving the coordinates of a location on the Earth measured in degrees latitude and East longitude. The Representation Information will typically include the definition of ASCII together with descriptions of the format of the numbers and their locations in the file, their definitions as latitude and longitude, and the definition of their units as degrees."

Let us start with the first sentence: "the bits represent an ASCII table of numbers." From our point of view the model is either conflating two different entities in one or the information as propositional content is not represented at all: a table of numbers is not information, it is a symbol structure, that expresses a particular kind of propositional content, much like our array of numerals expresses the number of students enrolled in graduate programs. The bit sequence *encodes* that symbol structure, the symbol structure that expresses propositions about latitude and longitude. The second sentence seems to confirm that the representation information for the bits includes both [1] the ASCII definition and [2] the definition of latitude and longitude. Thus the knowledge to "make sense" of a table of numerals is assigned as Representation Information for a set of bits, conflating in this case the expressing symbol structure — the table of numerals — with the encoding symbol structure, the bit sequence — that serializes the table.

The intricacy of the hierarchy of symbol-structure encodings — e.g., numerals encoded as ASCII characters, encoded as bits — has been investigated using the notion of Interpretive Frames (Dubin et al., 2011). Without a proper decoupling of these entities the identification of significant properties for datasets could not be performed at the correct level of abstraction.

### The PLANETS Conceptual Data Model

The relation between the PLANETS Conceptual Data Model with the OAIS Reference Model is stated explicitly in the specification (Sharpe, 2009):

> "...the model is designed to be compatible and to extend OAIS (e.g., explicitly defining different types of Information Object in need of preservation actions)."

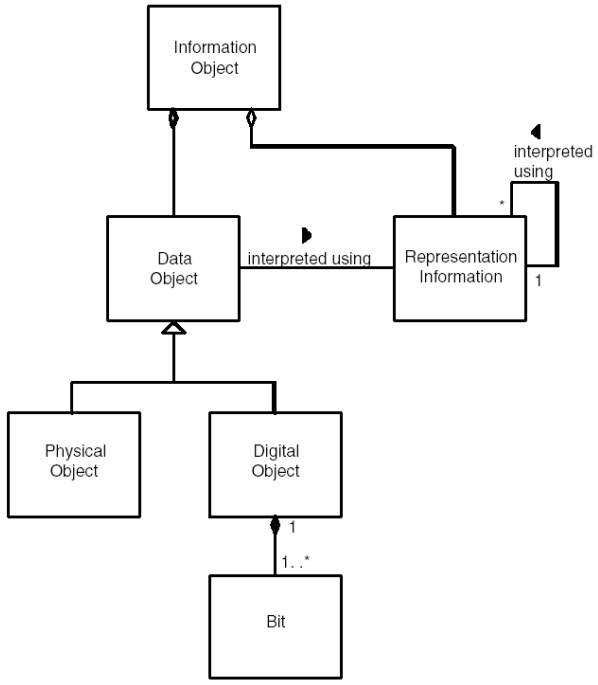


**Figure 3. OAIS Information Model.**

The PLANETS Conceptual Data Model is more detailed than the OAIS Information Model. While the PLANETS model still remains agnostic with respect to possible technological environments, its Conceptual Data Model component is more grounded in the technological aspects of the representation of digital objects in real information systems:

> "The main purpose of the model is to describe the actual preservation objects that need to be preserved. To enable preservation it is also necessary to model a variety of things such as formats, software, hardware, tools and properties that this model assumes already exist in a Registry."

PLANETS provides four conceptual entities: *Deliverable Unit*, *Expression*, *Manifestation*, and *Manifestation File*, defining a Deliverable Unit as:

> "An Information Object stored in an OAIS Archive for the purpose of supplying the Content Information to the Designated Community in a single DIP".

The emphasis in PLANETS is on the delivered object, which is preserved to fulfill the need of a Designated Community. The Manifestation entity accommodates the instantiation of the same Deliverable Unit by different Digital Objects. A Manifestation File is a logical representation of a file for a specific Manifestation and connects the conceptual model to the physical model that describes the actual files in a repository. Other conceptual entities called Components can be included at every level of the Conceptual Model to "help with the description of automatically measurable properties" for

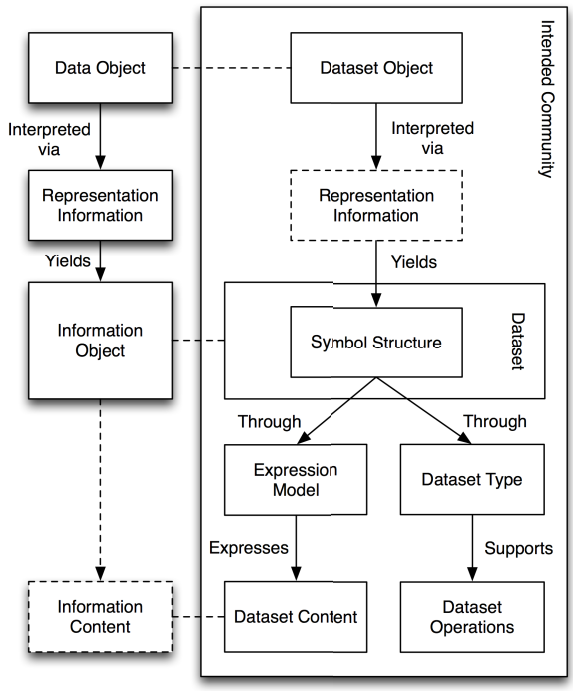


**Figure 4. Model alignment.**

each entity. Expressions are optional intermediate entities that fit between a Deliverable Unit and its Manifestations:

> "For example, an electronic journal could have three expressions for use by different audiences: a camera-ready expression (used for creating a printed version of the journal), a Web-based expression (used for publishing on-line) and an XML full text expression (used for search indexing)."

While the model provides many facilities for the identification of properties — in particular the Components allow for a fine-grained analysis of the properties of a Deliverable Unit, Expression, Manifestation, and Manifestation File — there is still conflation of propositional content and expression in the relation between an Information Object (a Deliverable Unit in PLANETS) and an Data Object (a Manifestation in PLANETS).

## MODEL ALIGNMENT AND SIGNIFICANT PROPERTIES

The InSPECT Project Framework Report (Knight et al., 2008) provides a more consistent interpretation of the OAIS model. The InSPECT Framework for Investigating Significant Properties of Electronic Content has been aligned with the OAIS Reference Model. We can align our conceptual model with the InSPECT Framework as a preliminary step for identifying significant properties of datasets.

The InSPECT Framework interprets the OAIS notion of Information Object as follows:

> "The Information Object has a central role within an OAIS, representing the product that must be recreated in order for a user to understand the information content."

This sentence assumes a difference between *Information Object* and *Information Content* that was treated inconsistently in the original OAIS Reference Model. Although Information Content is not officially recognized in OAIS, we feel that this distinction is a critical one and therefore follow the InSPECT interpretation in our alignment.

The InSPECT report refers to a "still image" as an example of Information Object in OAIS. The still image is "recreated" from a Data Object via its Representation Information: a Bitmap in a Submission Information Package, a TIFF in an Archival Information Package, and a JPEG in a Dissemination Information Package, are interpreted through their respective Representation Information to recreate the same "still image."

The "still image" in the InSPECT example appears to be at the same level of abstraction as our notion of Dataset. The information content of a still image is presumably a set of features a person can experience looking at the actual rendered image — the content expressed by the image. Similarly, Dataset Content is the information content expressed by a Dataset.

Following the InSPECT Framework's interpretation of the OAIS Information Model, we can align it with our Conceptual Model for Dataset as presented in Figure 4. The Dataset is an OAIS Information Object, with Dataset Content aligning to Information Content as described in the InSPECT Framework. A Dataset Object is interpreted in order to decode the sequence of bits as a symbol structure in a Dataset role. There are two kinds of information needed to fully interpret a Dataset: the Dataset Type gives the information needed to interpret the symbol structure from an operational point of view (it expresses structure and value type(s)), and the Expression Model gives the information needed to assign a symbol structure the role of bearing particular Dataset Content.

The interpretive aspects of the Expression Model and the Dataset Type are essential to scientific identity of a Dataset, since without them it will not be possible for a Dataset to be correct interpreted and used to support scientific claims. However, they are not fully accounted for in OAIS, where Representation Information is defined as:

> "The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol."

This notion of Representation Information seems to only account for the decoding of a bit-level Dataset Object into a another symbol structure (the Information Object in OAIS), and does not account for how the resulting symbol structure expresses information content or what operations can

be supported by the symbol structure.

The notion of interpretive frames (Dubin et al., 2011) may be able to provide a fine-grained comprehensive view of the how we go all the way from a Dataset Object to Dataset Content through the interpretation of a series of symbol structures. Further analysis will be need to give a full account of how OAIS Representation Information acts within interpretive frames in the decoding and use of datasets.

The PLANETS model could be aligned the same way for its compatibility with OAIS: the Deliverable Unit is aligned with our notion of Dataset, the Dataset Object with a Manifestation. Our model supplies an extension to the notion of Information Object in OAIS suitable for the representation of the entities involved in preserving a dataset. The OAIS Information Model itself is not operating at a level that provides us with the relevant entities for identifying datasets and determining whether preservation actions have been successful.

#### Signi cant Property categories for datasets

The InSPECT Report (Knight et al., 2008) defines Significant Properties with respect to the OAIS Reference Model, as follows:

> "Those characteristics of an information object that must be maintained to ensure that object's continued access, use, and meaning over time as it is moved to new technologies"

Significant properties are identified by an evaluator with respect to an Information Object for a specific community. In order to reflect the nature of datasets as symbol structures in particular roles, we need to identify and assign significant properties with respect to the entities relations involved in defining that role for some intended community. Significant properties of datasets then are not properties of the symbol structure by itself. Since it is the expressed Dataset Content and the supported Dataset Operations that qualify a symbol structure as a dataset, significant properties must be derived from those entities as well.

The draft revision of OAIS (2009) introduces the concept of Transformational Informational Property, defined as:

> "An Information Property whose preservation is regarded as being necessary but not sufficient to verify that the Non-Reversible Transformation has adequately preserved information content."

As noted by Giaretta et al. (2009), this concept is intended to correspond to the notion of a significant property.

A Reversible Transformation is defined in the OAIS Revision as:

> "A Transformation in which the new representation defines a set (or a subset) of resulting entities that are equivalent to the resulting entities defined by the original representation. This means that there is a one-to-one mapping back to the original representation and its set of base entities."

A Non-Reversible Transformation, on the other hand is:

> "A Transformation which cannot be guaranteed to be a Reversible Transformation."

Roughly speaking, a transformation is a migration between digital formats, and many preservation actions consist of these transformations. Giaretta (2009) argues that the notion of significant properties as defined above is not sufficient for application to scientific data. The notion of Transformational Informational Properties connects directly to the preservation actions that ensure access to digital objects and supports evaluation of those transformation-based methods.

We can apply the notion of Transformational Information Properties to our account of datasets by defining sub-categories that make reference to the entities in our model. This will move us closer to a fine-grained account of the successful preservation of scientific data.

As a preliminary step, we identified two sub-categories:

**Content properties:** those properties whose preservation is regarded as being necessary to verify that a transformation has adequately preserved the contingent relation between a symbol structure and the Dataset Content it expresses for the Intended Community. The properties are mostly to be derived from the Expression Model. This category can specialized. For example:

- Coordinate Systems Properties
- Temperature Value Systems Properties
- Column Semantics (for tables)
- Row Semantics (for tables)

**Operational properties:** those properties whose preservation is regarded as being necessary to verify that a transformation has adequately preserved the contingent relation between a symbol structure and the Dataset Operations it supports for the Intended Community. These properties are mostly to be derived from the Dataset Type. This category can be further specialized. Examples of specialization are:

- Data Structure Properties
- Data Value Properties
- Relational Kind Properties

Properties like the "The numeral tokens are of Cardinal Number type" are properties of the Symbol Sequence in a Dataset role. The identification and exploitation of these properties however depends on assigning them to the appropriate entity and not conflating entities or assigning properties to things that only have them in a derivative sense. This is what our model supports.

### Preservation of Signi cant Properties

As discussed at the start of this paper, the focus of interest in digital preservation are abstractions that do not literally undergo changes of state. How then do we account for success or failure in "preserving" properties of scientific datasets?

First, we recognize that significant properties of data may be instantiated at any level of abstraction. We may be concerned with a property of propositional content, such as its intensional relation to an entity in the domain of inquiry (e.g., "measuring wavelength"). Access to or correct interpretation of a dataset may depend on a notational or expressive property, such as "conforming to IEEE 754-2008." For some kinds of data, such as images and sound, physical properties of concrete exemplifying tokens may be significant ("being forty one seconds in duration" or "being eight centimeters wide," for example).

While abstract objects do not change, we recognize certain event types in what is colloquially termed the "information life cycle." But we characterize these as indication events, rather than changes to the state of an information resource. By indication we mean the selection or determination of an abstract pattern by an agent (Levinson, 1990). Indications include both ephemeral utterances and inscriptions: the fixing of a discrete symbol pattern in a tangible medium of expression. Writing notes on paper and saving a digital file to disk are both examples of inscriptions.

Inscriptions and other indication events typically employ mappings across different expressive levels, such as from bit sequences to EBCDIC character strings, or from EBCDIC characters to hole patterns in Hollerith cards. Some mappings are known and available to agents of preservation transactions, while others are unknown. For example, the standard mapping from UTF-8 encoded octet sequences to UCS character sequences is known, but mappings that can correctly govern interpretations of the Voynich Manuscript are currently unknown, may never be known, and might not exist.

Physical objects, unlike abstractions, do undergo real change. At any time a conformance relation obtains between a quantity of energy or matter and whatever its physical arrangement happens to be. That conformance relation can cease to obtain if the matter or energy is rearranged into some other pattern. So any physical object serving as the medium of a prior inscription event must maintain its arrangement in order to preserve a meaningful pattern. Now we can understand the preservation of a digital resource property to require:

- The existence of a physical medium that conforms to a prior inscribed pattern.

- A known mapping from the inscribed pattern to an abstract structure at a level directly indicated by an agent. This mapping might cross several expressive levels via function composition (bits to characters, characters to XML elements, etc.).

- Direct instantiation of the significant property by either the indicated structure itself (e.g., well-formedness), by an abstract object expressed or represented by that structure (e.g., falsity, "being greater than eleven") or by a concrete token of that abstraction ("being red" or "being thirty seconds long"). In the second and third cases we understand the indicated structure to "encode" (rather than instantiate) the property.

Several categories of preservation failures can be situated with respect to this model, such as:

- Scenarios in which storage media undergo physical destruction or damage. In these, the medium ceases to conform to the arrangement of its inscription event.

- Scenarios in which legacy data encoding formats are supported over time by fewer software applications. This can be seen as a loss of knowledge of one or more mappings composed to connect the medium's physical arrangement with the abstract object that instantiates or encodes the property of interest.

- Migration scenarios in which the resulting resource expression lacks a property instantiated by earlier expressions. Here the migration event indicates an abstraction that neither instantiates nor encodes the property.

## CONCLUSION

We have shown that when contemporary preservation models are revised to reflect an explicit and exact identification of entities and their relationships, distinguishing types of objects from the roles those objects enter into, and avoiding misleading idioms, the concept of significant properties does indeed have a promising application to scientific datasets.

## ACKNOWLEDGMENTS

The research reported here is being carried out at the Center for Informatics Research in Science and Scholarship (CIRSS) at the University of Illinois at Urbana-Champaign, Carole L. Palmer, Director. It is funded by the National Science Foundation as part of the Data Conservancy, a multi- institutional NSF funded project (OCI/ITR-DataNet 0830976) hosted at Johns Hopkins University Sheridan Libraries.

### References

Brown, A. (2008). Developing practical approaches to active preservation. *International Journal of Digital Curation*, *2*(1).

Cheney, J., Lagoze, C., & Botticelli, P. (2001). Towards a theory of information preservation. *Research and Advanced Technology for Digital Libraries*, 340–351.

Cox, S. (2006, September). *Observations and measurements* (OGC Best Practices document No. OGC 05-087r4). Open Geospatial Consortium Inc.

Dappert, A., & Farquhar, A. (2009). Significance is in the eye of the stakeholder. In *Proceedings of the 13th european conference on research and advanced technology for digital libraries* (pp. 297–308). Berlin, Heidelberg: Springer-Verlag. (ACM ID: 1812838)

Dubin, D. (2010, October). Encoded descriptions at face value. In A. Grove (Ed.), *Proceedings of the 73rd annual meeting of the american society for information science and technology* (Vol. 47). Pittsburgh, PA: Information Today, Inc.

Dubin, D., Futrelle, J., Plutchak, J., & Eke, J. (2009). Preserving meaning, not just objects: semantics and digital preservation. *Library Trends*, *57*(3), 595-610.

Dubin, D., Wickett, K. M., & Sacchi, S. (2011, August). Content, format, and interpretation. In B. T. Usdin (Ed.), *Proceedings of balisage: The markup conference.* Montreal, Quebec.

Farquhar, A., & Hockx-Yu, H. (2008, July). Planets: integrated services for digital preservation. *Serials: the Journal for the Serials Community*, *21*(2), 140–145.

Flouris, G., & Meghini, C. (2007). Some preliminary ideas towards a theory of digital preservation. In *Proceedings of the 1st international workshop on digital libraries foundations.*

Giaretta, D., Matthews, B. M., Bicarregui, J. C., Lambert, S. C., Guercio, M., Michetti, G., et al. (2009, October). Significant properties, authenticity, provenance, representation information and OAIS. In *Proceedings of iPRES 2009: the sixth international conference on preservation of digital objects* (pp. 67–73). San Francisco: University of California.

Guarino, N., & Welty, C. (2000). A formal ontology of properties. *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, 191–230.

Haugeland, J. (1981). Analog and analog. *Philosophical Topics*, *12*(1), 213–225.

Hedstrom, M., & Lee, C. A. (2002). Significant properties of digital objects: definitions, applications, implications. In *Proceedings of the DLM-Forum* (pp. 218–27).

Hockx-Yu, H., & Knight, G. (2008). What to preserve?: significant properties of digital objects. *International Journal of Digital Curation*, *3*(1).

Hourclé, J. A. (2008). FRBR applied to scientific data. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1–4.

IFLA Study Group on the Functional Requirements for Bibliographic Records (Ed.). (1998). *Functional requirements for bibliographic records: Final report* (Vol. 19). München: K. G. Saur.

Knight, G., Grace, S., & Montague, L. (2008, May). *Framework for the de nition of signi cant properties* (project document). London: InSPECT Project.

Kuhn, W. (2009). A functional ontology of observation and measurement. *GeoSpatial Semantics*, 26–43.

Levinson, J. (1990). What a musical work is. In *Music, art, and metaphysics: Essays in philosophical aesthetics* (pp. 63–88). Ithaca, NY: Cornell University Press.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., & Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, *2*(3), 279–296.

Matthews, B., McIlwrath, B., Giaretta, D., & Conway, E. (2008). The significant properties of software: A study. *JISC report, March*.

McDonough, J. P. (2011). Packaging videogames for long-term preservation: Integrating FRBR and the OAIS reference model. *Journal of the American Society for Information Science and Technology*.

Mois, M., Klas, C. P., & Hemmje, M. L. (2009). Digital preservation as communication with the future. In *Digital signal processing, 2009 16th international conference on* (pp. 1–8).

Moreau, L., Freire, J., Futrelle, J., McGrath, R., Myers, J., & Paulson, P. (2008). The open provenance model: An overview. *Provenance and Annotation of Data and Processes*, 323–326.

Reference model for an open archival information system (OAIS) [Computer software manual]. (2002, January). Washington, DC.

Reference model for an open archival information system (OAIS) [Computer software manual]. (2009, August). Washington, DC. (Draft Recommended Standard)

Renear, A. H., Sacchi, S., & Wickett, K. M. (2010, October). Definitions of dataset in the scientific and technical literature. In A. Grove (Ed.), *Proceedings of the 73rd annual meeting of the american society for information science and technology* (Vol. 47). Pittsburgh, PA: Information Today, Inc.

Renear, A. H., & Wickett, K. M. (2009). Documents cannot be edited. In *Proceedings of balisage: The markup conference* (Vol. 3).

Sandore, B., & Unsworth, J. (2010, June). ECHO DEPository—phase 2: 2008–2010 final report of project activities [section]. In (pp. 33–37). Champaign, IL: University of Illinois.

Sharpe, R. (2009). *PLANETS data model overview* (Tech. Rep. No. IF8-D1). (please request from info@planets-project.eu)

Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. *The state of digital preservation: an international perspective*, 2425.