Digital preservation and *the flow of information* in a digital environment

Towards sound conceptual and modeling foundations for digital preservation

Simone Sacchi — GSLIS Doctoral Candidate
sacchi1@illinois.edu

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois
ILLINOIS

This poster presents the first component of a formal framework for digital preservation. Digital resources are modeled from a cognitive perspective that addresses their communication function as information–carrying entities. Preservation expectation are expressed according to four basic units of analysis.

1. Introduction

This poster presents the first component of a modeling framework for digital preservation and a representative application scenario.

The framework

This framework focuses on how information flows in a digital environment.

> ***Digital preservation can be understood as a form of communication sustaining the flow of intended information from the creator of a digital resource to its potential users.***

The modeling paradigm is informed by works in Situation Theory and Situation Semantics [Devlin, 1995] and built on:

- Insights from previous influential models developed within the digital preservation community —in particular, the OAIS Reference Model [CCSDS, 2012] and the NAA Performance Model [Heslop, 2002].
- Conceptual work from the Data Concept group at Illinois —in particular the Basic Representation Model [Wichett et. al., 2012].

An application scenario

In the practice of digital preservation we face a high degree of heterogeneity in preservation expectation and strategies to address them. This is not surprising, given the variety of resources, scholarly practices, and communities involved.

When applying this framework common patterns emerge, allowing to analyze and expresses different preservation expectations according to more basic primitives. Three of these primitives emerge directly from the model: *Intended Information*, *Intended Representation*, *Intended Performance*. A forth one follows from them: *Artefact* —i.e. a resource along with the environment necessary to support its performance as originally conceived.

2. Situations Theory and Situation Semantics

Situation Theory (ST) and Situation Semantics (SS) together provide a logic–based mathematical framework to analyze information flows.

> ***For a cognitive agent information arises from situations that embody some form of representation of that information.***

On this account:

- A ***situation*** is a temporally and spatially located structured part of reality.
- A ***representation*** is a symbol structure that contingently expresses some information.

In particular:

> ***What information arises from a situation for an agent is a function of the agent's scheme of individuation —its ability to discriminate representations— and the constraints the agent is attuned to —that assign meaning to representations***

On this account:

- A ***scheme of individuation*** is the innate or acquired capability of cognitive agents to "carve up reality into cognitively manageable pieces" [Devlin, 1995].
- A ***constraint*** is a link between a situation carrying information —a *discourse situation*— and a situation the carried information is about — a *described situation*. Examples of constraints are "Natural laws, social conventions, linguistic conventions, etc." [Devlin, 1995].

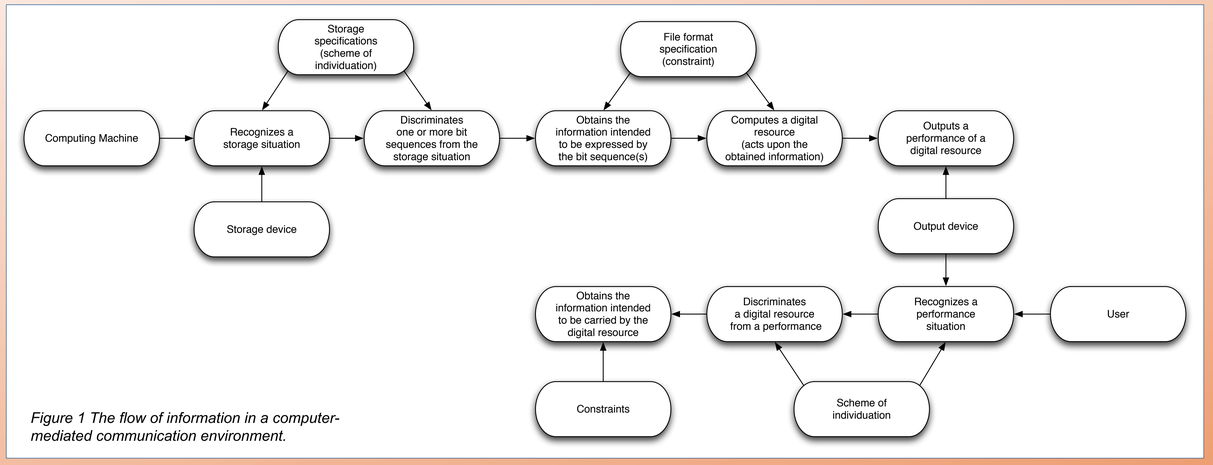

Figure 1 The flow of information in a computer-mediated communication environment.

3. Insight and issues from previous models

The **OAIS Reference Model** recognizes that information is obtained by applying an *interpretive frame* [Dubin et al., 2011, Sacchi et al., 2011] to a *symbol structure* that functions as a representation of that information.

> ***A Data Object interpreted via its Representation Information yields an Information Object***

However:

- OAIS does not explicitly model the fact that the representation meant for users' consumption —a *documentary form* [Duranti, 2005]— is different from the representation for storage —a *storage form*.

The **NAA Performance Model** recognizes the cognitive aspect of obtaining information in a computer–mediated environment. In particular the fact users obtain information from concrete embodiments of digital resources in the form of performances of those resources.

> ***When a source and a process interact a performance of a digital resource is produced***

However:

- The NAA Performance model does not account for how and what information arises from these performances.

The **Basic Representation Model** provides an ontologically precise account of how information is contingently represented in a digital environment in terms of *Content*, *Symbol Structures* and *Patterned Matter and Energy* [Wickett et al., 2012].

> ***Some content is contingently expressed by a symbol structure that, in turn, is encoded by other Symbol Structures, one of which is inscribed in some PME***

However:

- BRM does not address how these contingent relationship are established and sustained.

4. The flow of information in a digital environment

Specialized notions of *situation* and *representation* from ST and SS provide the machinery to model the flow od information in a digital environment (**Figure 1**):

> ***A user obtains information from a performance (a situation) that embodies a digital resource (a representation)***

- A **digital resource** is an *abstract symbol structure*, a state of affairs that is realized by its *performances*. Examples of digital resources are: a digital text, a digital record, a digital image, a digital game, etc.
- **Digital**, in this sense, means that these information–carrying resources require a computer environment to be accessed, stressing the contingent relationship between a particular resource and any bit level representation.

> ***The information a digital resource is intended to carry is function of constraints defined at the time of its creation***

- These constraints provide the intended semantics and assign meaning to the resource. A user must be attuned, or at least aware of, the appropriate constraints in order to obtain the intended information.

Consider a typical example from the sciences: *a digital dataset intended to provide information about temperatures in a region for a period of time*:

- A **digital dataset** is a *symbolic representation* (e.g. in the form of a table of numerals) a researcher recognizes, for example, form a rendering on screen
- A researcher attuned to the intended constraints is able to establish a link between a discourse situation —e.g. what he sees on screen— and a described situation —namely a spatio–temporal region— and obtain some information about it —namely, temperature values.
- In this case, the researchers must be attuned to constraints like: numeric conventions, units of measurements, date and time conventions, geographical conventions and so on.

ST and SS also provide the machinery to understand, in terms of information flow, how performances are produced from storage level representations:

- A **computing machine** discriminates *bit sequences* (storage representations) from a storage situation.
- According **to file format specifications** (constraints), it obtains for the bits the information necessary to compute a *digital resource*.
- A **digital resource** is then realized by a **performance** rendered on an output device.

The diagram in **Figure 2** represents as a formal ontology *entity types* and *relationship types* involved in all these processes.

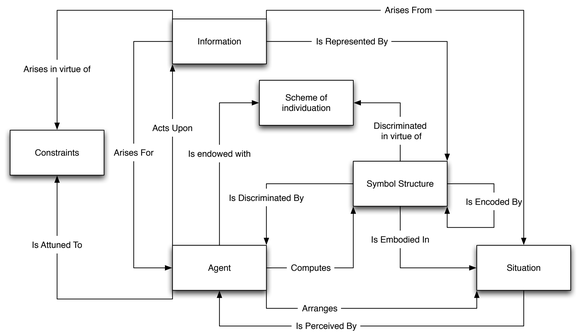

Figure 2: entity types and relationship types.

5. Units of analysis for digital preservation

Digital Preservation is ultimately about sustaining the flow of information from the creator of a resource to its potential users. However, different types of resources are used differently in different context and by different communities, leading to a variety of preservation expectations. The framework presented here allows for expressing complex expectation in terms of precisely defined basic primitives.

Intended information

The information a digital resource is intended to convey.

- **Conditions**: given the appropriate user constrains, any performance of a suitable representation should suffice.
- **Scenarios**: those where the preservation expectation is fulfilled by being able to access some content, regardless of the format. File format migration, data format migration, and emulation strategies are admissible.

Intended representation

A particular symbol structure that support certain information processing operations on the intended information it expresses.

- **Conditions**: given the appropriate user constraints and file format specification, any storage level representation should suffice.
- **Scenarios**: those where preservation expectations are fulfilled when particular types of data processing can be conducted. File format migration, and emulation strategies are admissible.

Intended performance (type)

A rendering of a resource conveying a particular sensory experience.

- **Conditions**: given file formats that include appropriate rendering constraint, any storage level representation should suffice.
- **Scenarios**: those where the preservation expectations involve particular visual or hearing characteristics. File format migration and emulation strategies are admissible.

Artefact

A resource in its original encoding and possibly the environment originally intended for its access

- **Conditions**: the original bit sequence encoding a digital resource plus the constraints to produce proper performances of that resource; possibly an intended hardware/software configuration.
- **Scenarios**: those where preservation expectations involve experiencing a resource as originally conceived tout court (e.g. digital art), those where file format conversion and media conversion are not admissible. Emulation strategies might be admissible.

References

CCSDS (2012) Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book).
Devlin, K. J. (1995). Logic and information. Cambridge University Press.
Duranti, L. (2005). The InterPARES Project: the long-term preservation of authentic electronic records: the findings of the InterPARES Project.
Heslop, H., Davis, S., & Wilson, A. (2002). An approach to the preservation of digital records. Canberra: National Archives of Australia.
Wickett, K. M., Sacchi, S., Dubin, D., & Renear, A. H. (2012). Identifying content and levels of representation in scientific data. Proceedings of the ASIS&T 75th Annual Meeting, 49(1), 1-10.

Acknowledgments

The work presented here had been developed as part of my dissertation research. I wish to thank my dissertation committee for their support and, in particular, Allen Renear, my academic advisor for the continuous help and inspiration.