Defining Digital Archaeology

Sergio Gregorio; Imaging & Media Lab, University of Basel & Swiss Federal Chancellery, Terminology Section (Switzerland)

Abstract

The following paragraphs examine the appropriateness of the expression digital archaeology as it is typically understood by the digital preservation community, discuss conceptual differences between possible meanings and propose an alternative designation (a label) for 'digital archaeology'. A brief description of terminological methods and applications may help sensitize the community towards the role terminology can play in avoiding linguistic inconsistencies and as an economic asset that provides added value to sustainable digital archives.

Introduction

Terminology is attracting interest in the digital domain as an increasingly indispensable component of information systems. Modern retrieval systems have come to rely on terminological knowledge in order to deliver quality search results. Digital asset management solutions and digital preservation in general contain and implicitly deal with a wealth of terminological concepts. Yet human creativity often leads to original linguistic coinage, superimposing known designations (words and expressions) on new and even on already familiar concepts. Unfortunately, this highly creative process leads to insufficient precision, which is detrimental to the exchange of scientific knowledge and ultimately to mutual understanding.

Archaeology and Natural Language

Archaeology proper, the science of antiquity, studying ancient cultures and trying to understand the past, relies on the remains of buildings, on physical artefacts that may have come to light by chance, and on inscriptions on different materials (stone, clay, papyrus) that have made it through the centuries. Thanks to these relatively stable physical media, it has been possible to hand down to posterity information about the past.

The code used for transmission was natural language, which in technical terms is inefficient, highly polysemous and thus leaves room for several interpretations. However, natural language offers an important intrinsic advantage over artificially codified languages that guarantees transmission under unfavourable communicative conditions: redundancy, and also the fact that we still understand ancient languages. Furthermore, this code might be described as 'open source', i.e. is shared by a large community of speakers. Thus, apart from having command of a specific language, no further effort is required for a language community to understand a message.

Machines and Digital Code

In the digital world there are no tangible remains or artefacts. Computers and accessory gadgets become obsolete and useless in a few years' time. It is hard to imagine a future archaeologist digging in settled layers of electronic waste, carefully brushing off dust from decades-old hard disks, tapes etc., and in doing so making unexpected and groundbreaking discoveries about our technological past. For the information our future archaeologist might be looking for is 'inscribed' on these data carriers, which need electricity to be operated, depend on a technology-oriented society and on the availability of electric power. If by chance these data carriers still happen to work, the next problem might be caused by incompatible cabling. But assuming that our archaeologist lives in the best of all worlds and manages to read the message on an ancient digital data carrier, the final and fundamental step will be the interpretation (decoding) of the digitally encoded information.

There are at least two mandatory conditions for the successful interpretation, viz. reconstruction of a digitally encoded message:

- The code (or language) must be known and a working decoder (computer programme or human expert) must be present
- The degree of damage must remain within limits in order to allow a sensible reconstruction using available redundant information.

But often the effort invested in terms of time and money does not produce the desired results. The recovered information reveals itself as useless or irrelevant. This is an exceedingly practical issue. A real-life example may clarify this point:

An Inquiry

E-Mail from August 31, 2005 (What follows is a translation which conveys the general sense of the original German version of the message):

"Dear Mr Gschwind, in 1991 you produced digitally colour reconstructed 35mm slides from 6x6 inch colour slides (from 1955) for the Cultures Museum (formerly Ethnology Museum) in Basel. Mr (…) had commissioned the work. Mr (…) is interested in documenting a before and after example of this reconstruction. The object in question is the 6x6 colour slide OZ 2253a.

Are you still in possession of the scanned versions of this slide made before and after the restoration and would it be possible to deliver the two files on a CD-ROM?"

Solution

- The tape still existed (DAT), but was no longer compatible with current tape drives (tape DDS1, current drive DDS4; tape was ejected after a few seconds)
- A DDS1 drive was sought, found and bought on Ebay. It was made usable after a laborious installation procedure
- Next problem: old SCSI I connection. SCSI I card and cable had to be bought (these had been disposed of years ago!).
- The read-back process was carried out successfully. The tape was undamaged and luckily the data had been written in the TAR-format
- *However,* the information was encoded in an "old" and proprietary data format. TIFF did not exist in 1991!

- This necessitated the search for and retrieval of the old Fortran programmes used to process the proprietary image files
- A conversion routine was written in C language to convert the proprietary image format to TIFF

8 months had elapsed since the inquiry! It took 3 days to deal with the inquiry, and to restore and process the data! The best part was still to come: the slide the museum was looking for had never been digitized! [1]

This is an example of the best of all worlds, in the sense that obsolete equipment could be put into operation again, and the data recovery process was carried out successfully with no data loss. But, more importantly, the person who implemented the old Fortran routines, the proprietary data format and was responsible for the project at that time was still working in the same office and so no drain of know-how had occurred, since without knowledge or documentation, a given bit-stream can represent almost anything [2].

Many further details could be adduced to underline the lucky circumstances under which it had been possible to retrieve decade-old data, among other things the fact that the institution that carried out the digitization and the colour reconstruction of the slides still had the data, even though it had not been entrusted with its custody.

Under these circumstances, a direct line to archaeology is hard to draw and a reference to archaeological methods appears somewhat far-fetched.

Digital Archaeology?

So, what is digital archaeology? Depending on the context, the expression may mean different things, for instance ground scanning methods to uncover the remains of ancient cultures, the use of digital equipment for the scientific analysis of artefacts and inscriptions or solving problems caused by the ageing of digital storage media and digital code, figuring out how to read stored bits and working out what they mean [3].

Since language usually lags behind technological innovation, linguistic quick fixes are often adopted to designate a new concept. Here there is the potential risk of coining words and expressions, which are eventually adopted by a community and gain term status through their extensive use. Digital archaeology in the latter sense is to be rated among these uses.

Terminology is the discipline that can help disentangle these meanings, taking apart and delimiting concepts, providing succinct definitions as well as assigning pertinent designations (words) to concepts. Terminology is attracting interest for various reasons, especially in the digital domain, and there is a rough understanding among specialists of all disciplines of what it is or should be.

Terminology

"Terminology is the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialized areas of usage of one or more languages" [4].

Terminology supports communication at all levels of interaction. In principle there is no difference between human information exchange in natural language or between information systems by means of digital code. Successful communication relies on recipient-appropriate information, which is in turn based on shared concepts.

Due to the complexity of natural language, concepts will, of course, be rarely clear-cut, but their specification in rough outlines will in any case be useful to systems and to users. This is also the reason for the increasing use of taxonomies, ontologies and classification systems as integrating knowledge components of information systems, where terminology plays a major role as an active element of information management.

Practical terminology work (terminography) is based on a few principles that make up its current methodology. In an extremely sketchy way these can be summarized as follows:

- Identification of terminological units (within specialized languages: What is a term? What object does it refer to?)
- Terminological definition (concise formulation of the delimiting characteristics of a given concept)
- Single-concept principle (one designation per concept, i.e. one dictionary entry per concept) [5]

The relationship between concept, object and designation (word, abbreviation, symbol etc.) is based on a simplified adaptation of the semiotic triangle [6]:

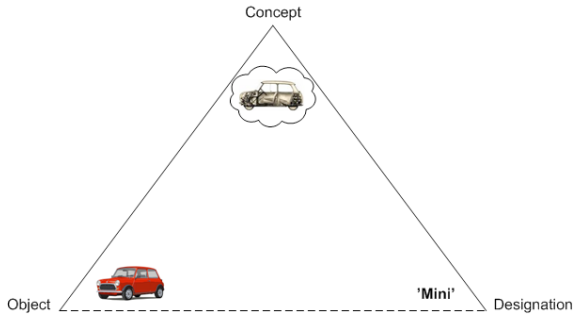

***Figure 1**. The "Mini"-Concept*

The concept conveys an idea and by means of characteristic features permits the identification of a specific object, which may be physically tangible, but can also be abstract or even an event. Of particular importance is the fact that there is no direct connection between the designation and the object (dotted line). The object is ultimately defined through its concept.

Concepts comprise the more or less specific characteristics of particular, individual objects (individual concepts; e.g. 'Mini') or whole classes of objects (general concepts, e.g. 'subcompact car'). These characteristics are used to define and delimit the concept by means of a definition and determine the position of the concept in a system of concepts, i.e. in a classification system [7].

Thinking in this framework and adopting the above mentioned methodology usually facilitates the choice of the proper designation among so-called 'competing labels' [8], especially in cases where different concepts 'share' the same designation, paving the way for polysemous usage and concepts becoming blurred and vague.

Thus, taking into consideration these principles and the above mentioned example of a successful resuscitation of information from what otherwise would have been a data and equipment

cemetery with the passage of time, and acknowledging that there might be a faint link to archaeology, an initial conclusion that can be drawn is that *digital epigraphy* appears to be the more appropriate term for the act of making sense of outdated or fragmented digital code from the past. But this solution would tenaciously chain the concept to the domain of archaeology, with which it shares a certain affinity when it comes to dealing with the interpretation and reconstruction of information from the past, but differs in a series of aspects.

There is within the domain of information technology a general concept that exactly describes the procedure, i.e. *data recovery*. Compared to the more appealing *digital archaeology,* it is of course utterly prosaic, but it fulfils its communicative function describing the recovery process. *Information restoration*, in the sense of bringing back a former state, can be understood as the process and at the same time as its successful conclusion. *Data reconstruction* might be a further option stressing the effort to regain potentially lost data [9].

These are of course recommendations dictated by common sense that more precisely describe the concept and contribute to transparent communication. Equally important, and maybe more convincing these days than purely linguistic reasons, is the economic impact of well-tended terminology collections on increasingly tight budgets.

Economic aspects of Terminology

There are important economic aspects involved in terminology work and good reasons for allocating financial resources to such a budgetary item. Leaving aside common misunderstandings or widely varying interpretations of a legal text that may lead to costly litigation, the benefits of good terminology exceed by far the production costs, especially through its reuse.

The German software company SAP, for instance, is using this proactive approach for practical reasons in their software localization projects [10]. For multilingual countries like Switzerland, it is a mandatory requirement for the public sector and supports the production of parallel legal texts in the official national languages. Even the ISO Standardisation Body has started their ISO Concept Database with the aim of collecting their terminological 'disiecta membra', streamlining their terminology in terms of consistency and for reuse [11].

Projects and current work

There are many initiatives with a lot of terminological work being done. In most cases uncoordinated efforts end up as a glossary of terms on a project's website.

There are excellent examples that go beyond this corollary function, for instance the nowadays ubiquitously cited document on the OAIS reference model [12], in which the concepts and deviations from common usage are clarified at the very beginning. Further examples are the dictionary produced by the InterPARES2 project [13], which mainly focuses on records management terminology and, within the same domain, the excellent Glossary of Archival and Records Terminology by the Society of American Archivists [14]. These are examples of terminology work carried out by subject-field specialists, even though the adopted methodology corresponds only in part to current terminological practice.

This is by no means a drawback, since these authoritative sources support the production of new glossaries, for instance the records management terminology project led by the Terminology Section of the Swiss Federal Chancellery, defining and proposing equivalents in the national languages and English, taking into account the different archival traditions with the objective of proposing a multilingual national standard. The above mentioned positive effects may be achieved by other scientific communities through collaboration, with the advantage of tackling terminology issues in a principled way. There are excellent starting points, for instance Bradley's article *Defining digital sustainability* [8], emphasizing the current shift of focus from purely methodological issues to increasingly economic concerns.

Taking this initiative a step further, a closer collaboration between domain specialists and terminologists can help preclude from the beginning potentially misleading, albeit colourful expressions like *digital archaeology* to the advantage of consistent and transparent communication.

Conclusion and possible future work

Compared to archaeology, digital preservation is still in its infancy. To date, short product life-cycles and fragile storage media have forced the digital archiving community to concentrate its efforts on the preservation of digital information (bit-stream preservation). There are no artefacts or remains at hand that may help reconstruct specific messages and contexts. Digital longevity remains a major challenge, although solutions for longer-lasting and persistent storage media are under development, e.g. the microfilm-based PEVIAR [15]. In order to contain the proliferation of buzzwords and promote accurate definitions for existing and new terms, the digital archiving community is invited to cooperate with terminologists. Good terminologies help improve communication between different scientific communities, provide interchangeable quality metadata to information systems and support record keeping processes within public authorities. Collaboration in general has become an indispensable prerequisite for sustainable digital archives. Terminological collaboration can add further value to digital preservation efforts and help reduce linguistic inconsistencies. Used accurately and with clear objectives in mind terminology can leverage current and upcoming information exchange activities.

References

[1] R. Gschwind & S. Gregorio, Workshop Archivierung digitaler Bilddaten, Imaging & Media Lab, University of Basel (2004)

[2] Jeff Rothenberg, Ensuring the longevity of digital information (1999) - expanded version of Ensuring the longevity of digital documents, Scientific American 272:1, pp 42-47 (1995)

[3] Caroline R. Arms, Keeping Memory Alive: Practices for preserving Digital Content at the NDLP of the Library of Congress, RLG DigiNews Vol. 4, N. 3 (2000)

[4] Juan C. Sager, A practical course in Terminology Processing, John Benjamins B. V. (1990)

[5] S. Pavel & D. Nolet, Handbook of Terminology, Public Works and Government Services Canada (2001)

[6] Charles K. Ogden & I. A. Richards, The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism, 8th ed, New York: Harcourt, Brace &World (1946)

[7] Conference of Translation Services of European States (COTSOES), Recommendations for Terminology Work, Swiss Federal Chancellery (2002)
[8] Kevin Bradley, Defining Digital Sustainability, Library Trends Vol. 56, N. 1, pp. 148-163 (2007)
[9] Seamus Ross & Anne Gow, Digital Archaeology: Rescuing Neglected and Damaged Data Resources, JISC/NPO Study (1999)
[10] SAP Terminology Database, http://help.sap.com/content/additional/terminology/info_terminology.htm (2009)
[11] ISO, ISO Concept Database - Common Features, Draft for discussion, STD_DBs N56 (2008)
[12] Reference model for an Open Archival Information System (OAIS), Blue Book, CCSDS 650.0-B-1, Consultative Committee for Space Data Systems (2002)
[13] International Research on Permanent Authentic Records in Electronic Systems, PARES 2 Project Dictionary, http://www.interpares.org/ip2/ip2_terminology_db.cfm (2008)
[14] Richard Pearce-Moses, A Glossary of Archival and Records Terminology, Society of American Archivists, http://www.archivists.org/glossary/ (2005)
[15] Permanent Visual Archive (PEVIAR), Research Project at the Imaging & Media Lab, University of Basel, http://www.peviar.ch

Author Biography

Sergio Gregorio studied English Philology, Italian Linguistics and Computer Science at the University of Basel, where he received his Master's degree in 1994. From 2003 to 2008 he worked as a research assistant at the Imaging & Media Lab of the University of Basel. He currently works as a terminologist and as a project leader for a large-scale records management project within the Swiss Administration with minor academic involvement (E-Mail:sergio.gregorio@bk.admin.ch)