Efthimios C. Mavrikas M.Sc.
CILab Researcher, PhD Candidate [1],[2]
tim@ct.aegean.gr

Evangelia Kavakli PhD
CILab Manager, Lecturer [1]
kavakli@ct.aegean.gr

Prof. Nicolas Nicoloyannis
ERIC Director, Professor [2]
nicolas.nicoloyannis@univ-lyon2.fr

[1] Cultural Informatics Laboratory (CILab)
Department of Cultural Technology and Communication
University of the Aegean
Sapfous & Arionos Street
81100 Mytilene
Lesvos
Greece
Phone/Fax: (30-22510) 36612

[2] Equipe de Recherche en Ingénierie des Connaissances (ERIC)
Université Lumière-Lyon 2
Campus Porte des Alpes
Bâtiment L
5, Avenue Pierre Mendès-France
69676 Bron Cedex
France
Tel: (33-4) 78774492 Fax: (33-4) 78772375

The Story between the Lines: Exploring Online Distributed Cultural Heritage Document Collections using Ontology-based Methods

Summary
Cultural Heritage documents deal with objects/artifacts and the people that created, owned, used, or (re)discovered them. Their fates are intertwined in unique and complex stories forming a cumulative body of knowledge, often fragmented across large online document collections. While our collective memory has explicitly documented these stories, the heterogeneity and distribution of the available sources creates islands of information that can only be implicitly connected by a limited, expert audience. This paper presents a semantically consistent framework for the online presence of Cultural Heritage document collections, set upon a participatory centre stage and supported by a shared knowledge model, the CIDOC CRM ontology. In this framework, Cultural Heritage document contributors are peer-to-peer network nodes that benefit from: a schema-based network topology; a transparent,

self-organised, self-optimised network infrastructure; knowledge-rich document processing modules which analyse and classify each contribution, capture the notion of time and the unfolding of events spanning a single or multiple documents, and establish meaning connectivity over the entire collection. Overall, this framework assists a scholarly audience with the exploration of online distributed Cultural Heritage document collections, and offers an informed tap into the collective memory scattered therein.

Cultural Heritage documents deal with objects/artifacts and the people that created, owned, used, or (re)discovered them. Their fates are intertwined in unique and complex stories forming a cumulative body of knowledge, often fragmented across online document collections. While our collective memory has explicitly documented these stories, the heterogeneity of the available sources creates islands of information that can only be implicitly connected by a limited, expert audience.

This paper presents a semantically consistent framework for the online presence of Cultural Heritage document collections, set upon a participatory centre stage and supported by a shared knowledge model. In this framework, Cultural Heritage document contributors benefit from knowledge-rich document processing modules which analyse and classify each contribution, capture the notion of time and the unfolding of events spanning a single or multiple documents, and establish meaning connectivity over the entire collection. Overall, this framework assists a scholarly audience with the exploration of online Cultural Heritage document collections, and offers an informed tap into the collective memory scattered therein.

Writing a usable past.

Collective remembering is grounded in a narrative tradition defined in terms of schematic narrative templates [19]. Cultural Heritage discourse draws on these *cognitive instruments* [13] to grasp together information about objects/artifacts, people, events, motivations and their setting(s) to *actively create the past in a modernist identity space* [18]. The historical, archaeological, museological, and anthropological discursive codes are assertive, affirmative accounts of *existents* and *occurrents* [2] reflecting a rigorous analysis of evidence while remaining shaped by implicit theories and partiality; relating to material evidence and finding knowledge, taking sides and making knowledge. The social and academic baggage of Cultural Heritage discourse is *represented by ideas which are represented by words* [1]. Words and writing create interpretations of the past unique to the textual medium, an intellectual product which resides equally in the aboutness of the texts and their linguistic material.

The linguistic material of a text includes all explicitly manifest linguistic signalling devices, including ecological, morphological, syntactic, semantic and pragmatic constructs. The ordering of text segments, the locational cues and cue phrases, the use of pronouns and reference, the tense and aspect of verbs are devices that often signal a specific *communicative intention*. However, the overall *intentional structure* [10] of the text includes *just as much that left unwritten as that written* [18]. The aboutness of the text has to be looked for in the silenced concepts and meanings, as well as the challenging of explicit signalling devices. Tilley, citing Macherey, citing Nietzsche, reminds us to ask the *hinterfrage* question of the text:

When we are confronted with any manifestation which someone has permitted us to see, we may ask: what is it meant to conceal? What is it meant to draw our attention from? What prejudice does it seek to raise? And again, how far does the subtlety of the dissimulation go? [15]

and eloquently sums up the issue:

To read a text adequately is to rewrite it, to fill in those absences found in the text's margins and the spaces between the lines and the words. A science of the text does not leave it where it is but transforms it. [18]

The scope of online presence.

Does a tree falling in the forest make a sound if no one is there to hear it?

A Cultural Heritage document – a written object in the sense of *information-as-thing* [5] – can only impart knowledge by reaching and matching an audience. The reader is the *introducer to the connecting links* [3] between points in the discursive network created by the intentional structures found inside a document collection. Document collections are constructed by *information institutions as physical places where resources are selected, organised, preserved and accessed* [4] in support of a community of readers. Online document collections expand the supported range of community settings to include traditional information institutions such as libraries, museums, archives and schools, concurrently with classrooms, offices, laboratories, homes and public spaces. The expanded community of readers is provided with the means to access, create, store, search, retrieve and explore online document collections within a Cultural Heritage scripting space.

A Cultural Heritage scripting space is a hypertextual design that engages a community of readers in a nonlinear and transformative mode of interaction with a Cultural Heritage document collection. A Cultural Heritage scripting space is a nontoy application of Semantic Web technology and computational semantics, oriented towards an expectation-based approach to meaning.

Handling analysis and generation of text.

The architecture of a Cultural Heritage scripting space heavily draws on the *ontological semantics* natural language processing theory and methodology introduced by Nirenburg and Raskin, by placing a constructed knowledge model – the ontology – at the core of an information extraction method and a reasoning method about knowledge derived from text and associated reader interactivity; the ontology is complemented by an episode repository, a lexicon, an onomasticon and a set of processing modules for the intentional structure of text.

The episode repository stores an *episodic memory* [16] of knowledge about instances of objects/artifacts, people, events, motivations and their setting(s), indexed by corresponding ontology concepts and interrelated on temporal, causal and other properties. The systematic indexing of episodic knowledge enables case-based reasoning [17] and analogical inference [6].

The lexicon lists common terms typically explained in terms of ontology concepts and referred episodic knowledge, indexed by citation form and connected to synonyms and related terms in controlled vocabularies. The onomasticon lists proper names typically explained in terms of ontology concepts, common terms listed in the lexicon and referred episodic knowledge, indexed by citation form and connected to unique terms in controlled vocabularies.

Processing the intentional structure of text involves the modularisation of linguistic signalling device analysis and generation into a set of modules operating in a pipeline.

A comprehensive analyser consists of:

- a tokeniser treating ecological issues such as special characters and strings, numbers, symbols, word boundaries and differences in fonts, alphabets and encodings;
- a morphological analyser dealing with the separation of lexical and grammatical morphemes and establishing meanings in grammatical morphemes;
- a semantic analyser containing a set of submodules:
  - a lexical disambiguator selecting an appropriate word sense from the list of senses enumerated in a lexicon entry;
  - a script-based semantic dependency builder establishing meanings in clauses;
  - a script-based discourse-level dependency builder establishing meanings in entire documents;
  - a script-based module determining the style of writing;
  - a script-based module managing the background knowledge necessary for the understanding of document contents, especially reference and coreference;

    - a module constructing reader profiles by determining the parameters of a reading situation – temporal, causal and otherwise – tracking the attitude of readers toward document contents and establishing reader intentions.

The analysis inputs are documents contributed to an online document collection. The analysis outputs are sets of knowledge structures – episodic knowledge referenced by lexicon and onomasticon entries – acquired from text and associated reader profiles. The analysis outputs are inputs to generation.

A comprehensive generator consists of:

- a content specification module establishing meanings in text to be generated and containing a set of submodules:
    - an interactive function specification module deciding on information inclusion or exclusion based on reader interactivity;
    - a profile function specification module deciding on information inclusion or exclusion based on assumed reader prior knowledge;
- a text structure module organising meanings into sentences and clauses and ordering them;
- a lexical selection module resolving semantic dependencies and idiosyncratic relationships such as collocation;
- a syntactic structure selection module;
- a morphological realiser of words;
- a lineariser of clauses and words.

The generation outputs are abstracts of contributed documents, derived from an evaluation of knowledge structure sets acquired during analysis and associated reader interactivity. The generation outputs correspond to the state of a Cultural Heritage scripting space for specific readers in response to their interactivity and assumed prior knowledge.

The transformative processing of the intentional structure of text based on reader interactivity and assumed prior knowledge aims to emulate a *reading for abstracting* of online documents – *exploratory-to-retrieval reading*, followed by *responsive-to-inventive reading*, followed by *connective (value-to-meaning) reading* [7] – by:

1. building a semantic representation for each document;
2. carrying out selection, aggregation and generalisation operations on each semantic representation to create new representations, leveraging reader actions and profiles;
3. rendering each new representation in natural language. [12]

Overall, the transformative processing of the intentional structure of text is an active reading for information content and a passive reading for understanding, uncovering the discursive network of an online document collection and encouraging a community of readers to undertake its research by *actively reading for understanding* and *passively reading for information content*. [7]

Handling narratives and readers.

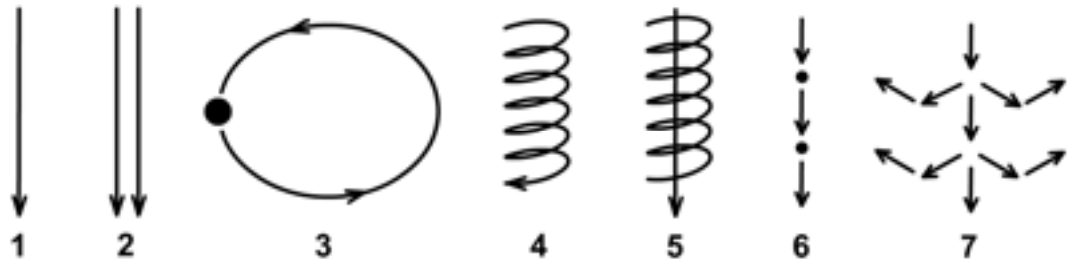
The expectation-based approach to meaning followed in a Cultural Heritage scripting space attempts to acquire and define schematic narrative templates in scripts explaining their cultural and cognitive constructions in terms of ontology concepts and referred episodic knowledge. Schematic narrative templates emerge out of narrative structures repeatedly identified in Cultural Heritage discourse and impose a basic plot structure on a range of objects/artifacts, people, events, motivations and their setting(s).

A schematic narrative template for historical discourse: *The Triumph over Alien Forces* [19] plays out as:
1. an initial situation, in which a nation is living in a peaceful setting where it presents no threat to others, disrupted by:
2. the initiation of trouble or aggression by alien forces, leading to:
3. a time of crisis and great suffering, which is:
4. overcome by the triumph over the alien forces by the nation, acting heroically and alone.

Other established schematic narrative templates for historical discourse include: *The Difficult Choice*, *The Quest for Freedom* [19] and *The Mystique of Manifest Destiny* [11].

A schematic narrative template for archaeological discourse: *The Detective Story* [18] plays out as:
1. a presentation of a problem to be solved, followed by:
2. an unfolding and discussion of putative clues, with the buildup of suspense until:
3. the tension is broken and the real meaning of the evidence is revealed.

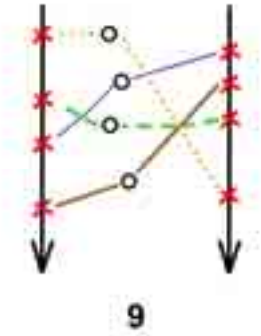
Scripts defining schematic narrative templates such as The Triumph over Alien Forces or The Detective Story, which emerge out of linear narrative structures (illustration 1), provide expectations for the processing of meanings of semantic and pragmatic constructs, the processing of style and the processing of reference and coreference.



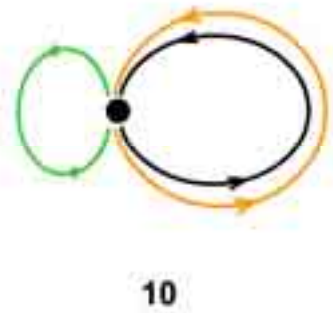
Scripts guide the generation of a hypertextual view of online document collections within a Cultural Heritage scripting space, by compiling document abstracts into alternative narrative structures, either noninteractive such as the:

- *parallel* (illustration 2), comparing different takes on a common theme side by side;
- *circular* (illustration 3), setting a central point as start and end to a theme exploration;
- *spiral* (illustration 4);
- linear with a *spiral mediation* (illustration 5);
- *aphoristic* (illustration 6), taking a point of departure and exploring a theme, stopping and starting off on an entirely different theme, never returning to the initial theme;
- *tangential* (illustration 7), taking a point of departure and exploring a theme, often exploring related themes at a tangent, returning periodically to the initial theme and exploring it in different directions; [18]

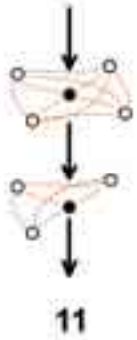
8



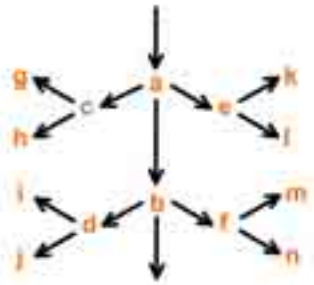
9

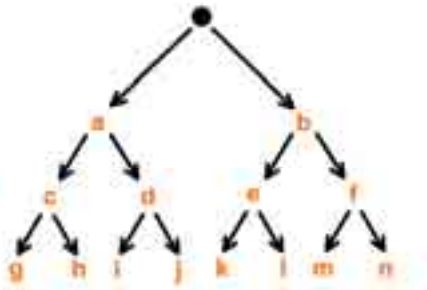


10



11





12

or interactive such as the:

- directed (illustration 8), prescribing a thematic itinerary with multiple exit and entry points to the main theme defined by the reader;
- braided (illustration 9), prescribing different takes on a common theme defined as a set of common concepts by the reader;
- orbital (illustration 10), prescribing a set of thematic explorations sharing a central point defined by the reader;
- shuffled (illustration 11), prescribing a sequence of thematic explorations with interchangeable departure points defined by the reader;
- arborescent (illustration 12), prescribing a main theme with tangential explorations defined as a hierarchy of decision points by the reader.

A tentative example.

A reader wishes to explore the concept of identity fragmentation and chooses Orientalism as a point of departure for this exploration. A series of interlinked abstracts are produced. The first abstract produces an Edward Said definition of Orientalism dating back to 1978, as quoted by Močnik. The second abstract produces a different passage by Močnik, contrasting Orientalism with Balkanism. The third abstract produces a passage by Ditchev, loosely referring to Balkanism.

According to Edward Said, Orientalism is a conceptual scheme, which ideologically mediates the domination of the West upon the Orient.
(...)
Contrary to Orientalism, where the logic of domination is imposed by colonial rule, in Balkanism, it is the immanent logic of self-constitution itself that generates the incapacity to conceive of oneself in other terms than from the point of view of the dominating other. [14]

In the case of semi-independent states like those in the Balkans, national actors are constantly torn between the need, on the one hand, to fit into the schemes of the geopolitical sponsors, abiding by general keywords, norms, and narratives, and the need on the other, to differentiate themselves and acquire an existence of their own in the universal imaginary of modernity. [9]

Key:
- Cue phrase
- Author, an onomasticon entry
- Representation, an onomasticon entry
- Phrase, an episode referring to a Representation; key constituent terms are lexicon entries
- Representation, an onomasticon entry
- Phrase, an episode referring to a Representation; key constituent terms are lexicon entries
- Phrase, an episode; key constituent terms are lexicon entries

---

The following pseudocode is a partial usage example of the CIDOC Conceptual Reference Model [8] as the knowledge model of a Cultural Heritage scripting space:

Onomasticon entry
|Orientalism|
Instance of E29 Design or Procedure
P69 is associated with:

Onomasticon entry
|Balkanism|
Instance of E29 Design or Procedure
P103 was intended for:
Onomasticon entry
|Representation|
Instance of E55 Type
P137 is exemplified by:
Episode
|a conceptual scheme, which ideologically mediates the domination of the West upon the Orient|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Phrase|
Instance of E55 Type
P72 has language:
Onomasticon entry
|English|
Instance of E56 Language
P94 was created by:
Lexicon entry
|definition|
Instance of E65 Creation Event
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
P4 has time-span:
Lexicon entry
|publication|
Instance of E52 Time-Span
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
P78 is identified by:
Episode
|1978|
Instance of E50 Date
P14 carried out by:
Onomasticon entry
|Edward Said|
Instance of E21 Person
P106 is composed of:
Lexicon entry
|concept|
Instance of E33 Linguistic Object

P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Lexicon entry
|scheme|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Lexicon entry
|ideology|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Lexicon entry
|mediation|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Lexicon entry
|domination|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
P129 is about:
Lexicon entry
|command|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Lexicon entry
|control|
Instance of E33 Linguistic Object
P2 has type:
Onomasticon entry
|Word|
Instance of E55 Type
Onomasticon entry

|West|
Instance of E53 Place
P87 is identified by:
Onomasticon entry
|Western Europe|
Instance of E44 Place Appellation
P70 is documented in:
Onomasticon entry
|UNESCO Thesaurus|
Instance of E32 Authority Document
P67 refers to:
Episode
|databases.unesco.org/thesaurus/|
Instance of E31 Document
Onomasticon entry
|Orient|
Instance of E53 Place
P87 is identified by:
Onomasticon entry
|Middle East|
Instance of E44 Place Appellation
P70 is documented in:
Onomasticon entry
|Getty Thesaurus of Geographic Names|
Instance of E32 Authority Document
P67 refers to:
Episode

|www.getty.edu/research/conducting_research/vocabularies/tgn/|
Instance of E31 Document
P102 has title:
Episode
|Middle East (general region)|
Instance of E35 Title
P1 is identified by:
Episode
|7001526|
Instance of E42 Object

Identifier
P106 forms part of:
Episode
|The Balkans as an Element in Ideological Mechanisms|
Instance of E73 Information Object
P128 is carried by:
Episode
|Balkan as Metaphor|
Instance of E84 Information Carrier

Acknowledgements.

This paper discusses research work undertaken by the first author in partial fulfilment of the requirements for the PhD degree of the University of the Aegean and the Université Lumière-Lyon 2, following a joint thesis supervision agreement between the two institutions (cotutelle de thèse). This research work is financed by the Hellenic Ministry of National Education and Religious Affairs and co-financed by the European Union.

Bibliography.

[1] Althusser, L. (1971) *Lenin and Philosophy and Other Essays*, London: New Left Books.
[2] Barthes, R. (1977) *Image, Music, Text*, New York NY: Hill & Wang.
[3] De Beer, C.S. (1996) *Ontology of the Book: Text, Intertext, Hypertext*, in South African Journal of Library and Information Science 15 (6): 124-129.
[4] Borgman, C. (2003) *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, Cambridge MA: MIT Press.
[5] Buckland, M.K. (1991) *Information as Thing*, in Journal of the American Society for Information Science 42 (5): 351-360.
[6] Carbonell J.G. (1983) *Learning by Analogy: Formulating and Generalizing Plans from Past Experience*, in *Machine Learning: An Artificial Intelligence Approach*, Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. editors, volume 1: 137-161.
[7] Cremmins, E.T. (1996) *The Art of Abstracting*, Arlington: Information Resources Press.
[8] Crofts, N., Doerr, M., Gill, T., Stead, S. and Stiff, M. editors (2004) *Definition of the CIDOC Conceptual Reference Model and Crossreference Manual v4.0*, ICOM/CIDOC CRM Special Interest Group.
[9] Ditchev, I. (2002) *The Eros of Identity*, in *Balkan as Metaphor: Between Globalization and Fragmentation*, Bjelić, D. and Savić, O. editors, Cambridge MA: MIT Press: 235-250.
[10] Kintsch, W. and Van Dijk, T.A. (1978) *Toward a Model of Text Comprehension and Production*, in Psychological Review 85 (5): 363-394.
[11] Lowenthal, D. (1985) *The Past is Another Country*, Cambridge: Cambridge University Press.
[12] Mani, I. (2001) *Automatic Summarization*, Amsterdam: John Benjamins.
[13] Mink, L.O. (1978) *Narrative Form as a Cognitive Instrument*, in *The Writing of History: Literary Form and Historical Understanding*, Canary, R.H. and Kozicki, H. editors, Madison WI: University of Wisconsin Press: 129-149.

[14] Močnik, R. (2002) *The Balkans as an Element in Ideological Mechanisms*, in *Balkan as Metaphor: Between Globalization and Fragmentation*, Bjelić, D. and Savić, O. editors, Cambridge MA: MIT Press: 79-115.
[15] Nietzsche, F. (1881) *Daybreak*, Cambridge: Cambridge University Press.
[16] Nirenburg, S. and Raskin, V. (2004) *Ontological Semantics*, Cambridge MA: MIT Press.
[17] Riesbeck, C. and Schank, R. (1989) *Inside Case-based Reasoning*, Hillsdale NJ: Lawrence Erlbaum.
[18] Tilley, C. (1990) *On Modernity and Archaeological Discourse*, in *Archaeology After Structuralism*, Bapty, I. and Yates, T. editors, London: Routledge. Also published online at archaeology.kiev.ua/meta/tilley.html
[19] Wertsch, J.V. (1998) *Mind as Action*, New York NY: Oxford University Press.