Science & Philosophy Volume 9(1), 2021, pp. 161-172

Epistemic logic and CERMINE: a logical model for automatic extraction of structured metadata

Simone Cuconato♦

Abstract

In this article we develop a logical model for the automatic extraction of structured metadata. We introduce a new predicate E – reads 'extract' – and a structure $\mathcal{S}$ to syntactically and semantically define metadata extracted with any automatic metadata extraction system. These systems will be considered, in the logical model created, as *knowledge extraction agents* (henceforth *KEA*). In this case *KEA* taken into consideration is CERMINE, a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form.

Keywords: epistemic logic; applied non-classical logics; logical methods in data science and knowledge engineering; metadata formalization; CERMINE

♦ Department of Informatics, Modeling, Electronics and Systems Engineering – DIMES, University of Calabria, Via P.Bucci, 87036, Rende (CS), Italy. simone.cuconato@unical.it

† Received on April 28th, 2021. Accepted on June 20th, 2021. Published on June 30th, 2021. doi: 10.23756/sp.v9i1.595. ISSN 2282-7757; eISSN 2282-7765.

.

1. A world of (meta)data

In the information world, at the most elementary level, metadata are defined as 'data about data' [11]. This basic definition is often detailed by referring to the structured nature of metadata and/or their machine-readable character. We need metadata in order to use the data to represent things that matter to us: to understand phenomena, to better serve customers, to establish organizational policies, and/or to keep a more accurate record of human activities. Within information systems, metadata perform a range of functions. These include:

- Searching: identifying the existence of a resource by keyword searching, browsing indexes or visualization techniques.
- Resource management: collection and database management.
- Selection: analysis and evaluation based on the description provided.
- Semantic interoperability: allowing searching across domains by means of equivalent elements.
- Location: finding a particular instance of a resource.
- Integrity and accountability verification and rights management
- Terms of availability information.

The Digital Library Federation identifies three types of metadata about digital resources:

- *descriptive metadata*: information describing the intellectual content of the object;
- *administrative metadata*: information necessary to allow a repository to manage the object;
- *structural metadata*: information that ties each object to others to make up logical units.

Metadata, whether descriptive, administrative or structural, ultimately share a single multifunctional goal: to contribute to a clearer and more modular management of digital objects and content retrieval. Automated metadata extraction enables the direct extraction of metadata from document sources. Obtaining structured metadata from documents, including title, authors, and publication date, is important to support retrieval tasks in information sciences. Various tools and frameworks exist to automatically extract this information

from PDF documents. Frameworks such as CERMINE, for example, are able to automatically extract metadata from specific document sources.

CERMINE [13, 14] is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form[1]. The system is based on a modular workflow and the implementations of most steps are based on supervised and unsupervised machine-learning techniques. The modular workflow, depicted in Figure 1.1 and 1.2, consists of three paths (*ii* and *iii* run in parallel): *i*) the base structure extraction path requires a pdf file as input and produces a geometric hierarchical structure in TrueViz format [6]. TrueViz is a tool capable of classifying the entities of each page of the structure into four categories: areas, lines, words and characters. In turn, each zone is labelled according to four other categories: metadata, references, body and other; *ii*) metadata extraction path analyses metadata parts of the geometric hierarchical structure. The result is a set of document's metadata from them in an XML format; *iii*) references extraction extracts a list of document's parsed bibliographic references.

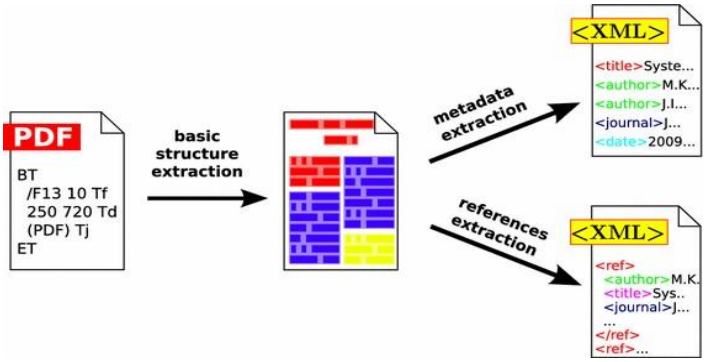

Figure 1.1: CERMINE's extraction workflow architecture[13]

[1] CERMINE system is available under an open-source licence and can be accessed at http://cermine.ceon.p.

Table 2 The decomposition of CERMINE's extraction workflow into independent processing paths and steps

Path	Step	Goal	Implementation
A. Basic structure extraction	A1. Character extraction	Extracting individual characters along with their page coordinates and dimensions from the input PDF file	iText library
	A2. Page segmentation	Constructing the document's geometric hierarchical structure containing (from the top level) pages, zones, lines, words and characters, along with their page coordinates and dimensions	Enhanced Docstrum
	A3. Reading order resolving	Determining the reading order for all structure elements	Bottom-up heuristic-based
	A4. Initial zone classification	Classifying the document's zones into four main categories: *metadata*, *body*, *references* and *other*	SVM
B. Metadata extraction	B1. Metadata zone classification	Classifying the document's zones into specific metadata classes	SVM
	B2. Metadata extraction	Extracting atomic metadata information from labelled zones	Simple rule-based
C. Bibliography extraction	C1. Reference strings extraction	Dividing the content of *references* zones into individual reference strings	K-means clustering
	C2. Reference parsing	Extracting metadata information from references strings	CRF

Figure 1.2: The decomposition of CERMINE's extraction workflow [14]

A system such as CERMINE can be formally represented through the use of epistemic logic. There is a close relationship between logic and computer science can be seen from the number of publications whose titles link the two disciplines with prepositions suggesting various degrees of proximity, cooperation or subordination: 'Logic in Computer Science', 'Logic and Computer Science', 'Logic for Computer Science', etc. However, some areas of computer science and technology, and in particular those relating to knowledge management and extraction, do not seem to have created the close relationship indicated above. The aim of this article is to create an innovative logic model applicable to engineering and data science [8].

2. Multi-agent epistemic logic: syntax and semantics

Epistemic logic [4, 15, 16] is the logic of knowledge and belief. Syntactically, the language of propositional epistemic logic is simply a matter of augmenting the language of propositional logic with a unary epistemic operator $K_i$such that

$K_i\varphi$ reads 'Agent i knows φ' for

some arbitrary proposition φ.

Hintikka provided a semantic interpretation of epistemic and belief operators which we can present in terms of standard possible world semantics along the following lines [7]:

> $K_i\varphi$: *in all possible worlds compatible with what i knows, it is the case that φ*

However, the general study of formal semantics for knowledge and belief (and their logic) really began to flourish in the 1990s with fundamental contributions from computer scientists [5, 10] and game theorists [3].

Definition 2.1 [Language] Let P be a set of atomic propositions, and A a set of agent-symbols. The language $\mathcal{L}_K$, the language for multi-agent epistemic logic, is generated by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_i\varphi$$

We use standard possible worlds semantics to give an interpretation to the language above. A model will be built on a set of epistemic alternatives (or worlds), and a relation built on these.

Definition 2.2 [Frames, Models, and Satisfaction] A Kripke Frame $F = (W, R)$ is a tuple where W is a set of epistemic alternatives for the agent, and $R \subseteq W \times W$ is an accessibility relation. A Kripke Model $M = (F, \pi)$, is a tuple where F is a Kripke frame and $\pi: P \to 2^W$ is an interpretation for a set of propositional variables P.

Given a model M and a formula φ, we say that φ is true in M at world w, written $M, w \vDash \varphi$ if:

- $M, w \vDash p$ iff $w \in \pi(P)$,
- $M, w \vDash \neg\varphi$ iff it is not the case that $M, w \vDash \varphi$,
- $M, w \vDash \varphi \wedge \psi$ iff $M, w \vDash \varphi$ and $M, w \vDash \psi$,
- $M, w \vDash K_i\varphi$ iff $(om\, w')(wRw'$ then $\mathcal{M}, w' \vDash \varphi)$.

A formula φ is valid, written $\vDash \varphi$, if it is true in every world in every model. I write $F, w \vDash \varphi$ to represent $M, w \vDash \varphi$ where M is an arbitrary model whose underlaying frame is F[2].

3. Knowledge formalization: syntax, semantics, axioms and structure

In order to deal adequately with *KEA* we extend the language $\mathcal{L}_K$ to obtain a language $\mathcal{L}_{K^+}$ The alphabet of $\mathcal{L}_{K^+}$ contains a new two-place predicate E such that:

E reads 'exstract'

Definition 3.1 [Language] The language $\mathcal{L}_{K^+}$ is generated by the following:

$$\psi ::= \Gamma^K_i p_{w_i} \mid \neg\psi \mid (\psi \wedge \psi)$$

where the intended interpretation of a formula $\Gamma^K_i p_{w_i}$ is 'agent i knows p_{w_i}', with p_{w_i} such that:

$$p_{w_i} =_{df} E^{m_i}_{d_i}$$

where $E^{m_i}_{d_i}$ reads 'extracts metadata m_i from document d_i'.

Definition 3.2 [Frames, Models, and Satisfaction] Given a Kripke Frame $F = (W, R)$, a Kripke Model $M = (F, \pi)$ and a formula ψ, we say that ψ is true in M at world w, written $M, w \vDash \psi$ if:

- $M, w \vDash \Gamma^K_i p_{w_i}$ iff $K_i\left(E^{m_i}_{d_i}\right)$,
- $M, w \vDash K_i p_{w_i}$ iff $(om\, w')(wRw' \text{then } \mathcal{M}, w' \vDash \psi)$,
- $M, w \vDash p_{w_i}$ iff $w \in \pi(P)$,
- $M, w \vDash \neg\psi$ iff it is not the case that $M, w \vDash \psi$,
- $M, w \vDash \psi \wedge \phi$ iff $M, w \vDash \psi$ and $M, w \vDash \phi$.

[2] We assume the standard definitions for metalogical properties such as axiomatisation, completeness, etc.

Definition 3.3 [Axioms and Inference Rules] The model on the language $\mathcal{L}_{K^+}$ is a first-order multi-modal version of the normal propositional system S5 and contains the following schemes of axioms and inference rules:

TAUT	Every classic propositional tautology
DIST	$K_i(\varphi \to \psi) \to (K_i\varphi \to K_i\psi)$
T	$K_i\varphi \to \varphi$
4	$K_i\varphi \to K_iK_i\varphi$
5	$\neg K_i\varphi \to K_i\neg K_i\varphi$
MP	$\varphi \to \psi, \varphi \Rightarrow \psi$
NEC	$\varphi \Rightarrow K_i\varphi$
EX	$\forall x\varphi \to \varphi[x/t]$
GEN	$\varphi \to \psi[x/t] \Rightarrow \varphi \to \forall x\psi$, x not free in φ
ID	$t = t$
FUNC	$t = t' \to (t''[x/t] = t''[x/t'])$
SUBST	$t = t' \to (\varphi[x/t] \to \varphi[x/t'])$

Definition 3.4 [Structure] Consider the following structure $\mathcal{S} = \langle A, W, P_{w_i}, M, D\rangle$:

- $A = \{a, b, c, \dots\}$ is a non-empty finite set of *KEA*,
- $W = \{w_1, \dots, w_m\}$ is a non-empty set of possible worlds ($|W| = m \in \mathbb{N}$),
- $P_{w_1} = \left\{p_{1_{w_1}}, \dots, p_{m_{w_m}}\right\}$ is a non-empty set of propositions $\left(\left|P_{w_1}\right| = m \in \mathbb{N}\right)$,
- $M = \{m_1, \dots, m_m\}$ is a non-empty set of metadata ($|M| = m \in \mathbb{N}$),
- $D = \{d_1, \dots, d_m\}$ is a non-empty set of documents ($|D| = m \in \mathbb{N}$) .

$\mathcal{S}$ is a dynamic structure, a structure in which possible worlds W occur. A is the set of *KEA*, while $P_{w_1}$is the set of epistemic propositions. M is the set of metadata while D is the set of documents.

4. Example of $\mathcal{S}$

Let us now consider two metadata extractions using the CERMINE system. The first extraction w_1 can be defined as a standard extraction as it was performed from born-digital scientific literature, specifically, extraction was

performed on two SARS-CoV-2 (covid-19) studies [9, 12]; on the contrary, the second extraction w_2 can be defined as a *non-standard extraction* since it was performed on a logic [1] and an information science [2] article.

Consider the following structure $\mathcal{S} = \langle A, W, P_{w_i}, M, D \rangle$:

- $A = \{c\}$;
- $W = \{w_1, w_2\}$;
- $P_{w_{1,2}} = \left\{p_{1_{w_1}}, \dots, p_{m_{w_1}}, p_{1_{w_2}}, \dots, p_{m_{w_2}}\right\}$
- $M = \{m_1, m_2, m_3, m_4\}$
- $D = \{d_1, d_2, d_3, d_4\}$

$P_{w_{1,2}}$:

$$\forall_i \in A, \psi_i: \begin{Bmatrix} K_c(E_{d_1}^{m_1}), K_c(E_{d_1}^{m_2}), K_c(E_{d_1}^{m_3}), K_c(E_{d_1}^{m_4}), \\ K_c(E_{d_2}^{m_1}), K_c(E_{d_2}^{m_2}), K_c(E_{d_2}^{m_3}), K_c(E_{d_2}^{m_4}), \\ K_c(E_{d_3}^{m_1}), K_c(E_{d_3}^{m_2}), K_c(E_{d_3}^{m_3}), K_c(E_{d_3}^{m_4}), \\ K_c(E_{d_4}^{m_1}), K_c(E_{d_4}^{m_2}), K_c(E_{d_4}^{m_3}), K_c(E_{d_4}^{m_4}) \end{Bmatrix}$$

$c = Cermine,\ m_1 = Title, m_2 = Abstract,\ m_3 = Author,$
$m_4 = Keywords$
$d_1 = [12], d_2 = [9]$
$d_3 = [1], d_4 = [2]$

$\boldsymbol{w_1}$:

- $K_c(E_{d_1}^{m_1}) = \mathrm{T}$
- $K_c(E_{d_1}^{m_2}) = \mathrm{T}$
- $K_c(E_{d_1}^{m_3}) = \mathrm{T}$
- $K_c(E_{d_1}^{m_4}) = \mathrm{T}$
- $K_c(E_{d_2}^{m_1}) = \mathrm{T}$
- $K_c(E_{d_2}^{m_2}) = \mathrm{T}$
- $K_c(E_{d_2}^{m_3}) = \mathrm{T}$
- $K_c(E_{d_2}^{m_4}) = \mathrm{F}$

In the first extraction w_1, agent *c* correctly extracts seven out of eight

metadata items. In the first paper *c* correctly extracts all metadata. In the second paper, however, it correctly extracts only three metadata: m_4, despite being present in the document, is not captured. This extraction can be represented of the model of Figure 4.1

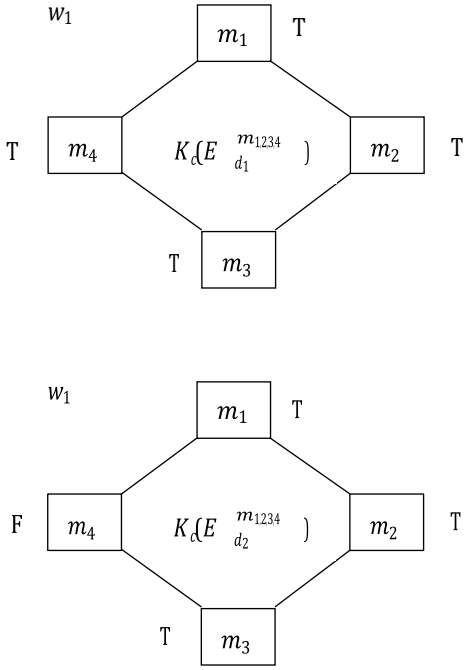

Figure 4.1: The model of $\mathcal{S}$ in w_1

$\boldsymbol{w_2}$:

- $K_c(E^{m1}_{d_3})$ = T
- $K_c(E^{m2}_{d_3})$ = T
- $K_c(E^{m3}_{d_3})$ = T
- $K_c(E^{m4}_{d_3})$ = F
- $K_c(E^{m1}_{d_4})$ = T
- $K_c(E^{m2}_{d_4})$ = T
- $K_c(E^{m3}_{d_4})$ = T
- $K_c(E^{m4}_{d_4})$ = T

In the second extraction w_2 the agent c correctly extracts seven out of eight metadata. In the third document *c* does not report any information about

metadata m_4, although the metadata is present within the article. In the fourth document c correctly extracts all metadata. This extraction can be represented of the model of Figure 4.2

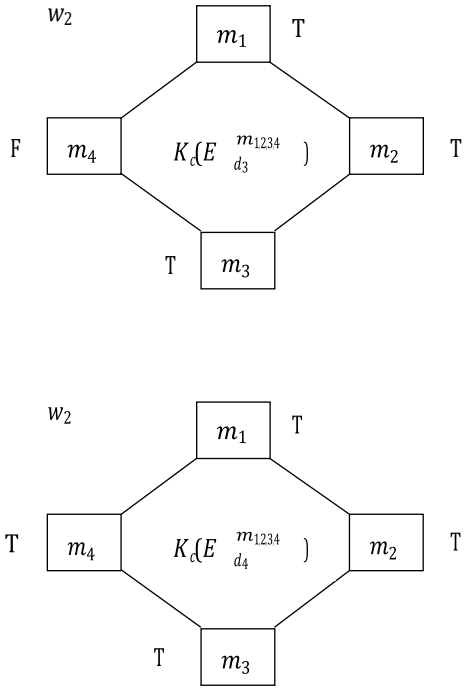

Figure 4.2: The model of $\mathcal{S}$ in w_2

5. Conclusions and future work

The creation of logic models applicable to automatic metadata extraction systems is a new but extremely interesting research topic. In this article we have provided a formalisation of the knowledge extracted by a specific extraction system: CERMINE. The use of formal models in support of data science and knowledge engineering offers many advantages to knowledge managers: *i*) it guarantees a rapid study of the extracted information; *ii*) it allows a comparison between different automatic metadata extraction systems; *iii*) it allows visualising the presence of inconsistent or incomplete metadata. The last point is probably the most challenging aspect of logic research applied to automatic

metadata extraction systems. A possible future development could be: *i*) to study the problem of inconsistent or incomplete metadata in a similar way to how it was treated in inconsistent or incomplete databases; *ii*) the use of polyvalent logics suitable for dealing with vague, imprecise or unreliable data.

References

1. F. Belardinelli, A. Lomuscio. "A quantified epistemic logic for reasoning about multiagent systems". *6th International Joint Conference on Autonomous Agents and Multiagent Systems*. 2007.
2. E. Bottazzi, R. Ferrario. "Preliminaries to a DOLCE Ontology of Organizations". *Int. J. Business Process Integration and Management*. Vol. 4, No. 4, 2009.
3. Bonanno and Battigalli. "Recent results on belief, knowledge and the epistemic foundations of game theory". *Research in Economics*. Volume 53, Issue 2, June 1999, Pages 149-225. 1999.
4. S. Cuconato. "A logical framework for democratic decision-making: epistemic logic and liquid democracy", *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 2020.
5. R. Fagin, J. Y. Halpern, Y. Moses, M. Y. Vardi. *Reasoning About Knowledge*. The MIT Press: Cambridge, MA. 1995.
6. C. Ha Lee, T. Kanungoa. "The architecture of TrueViz: a groundTRUth=metadata editing and VIsualiZing ToolKit", *Pattern Recognition*, 36 pp. 811 – 825. 2003.
7. J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Second edition, Vincent F. Hendriks and John Symons (eds.), (Texts in Philosophy, 1). London: College Publications. 1962 [2005].
8. H. Hosni, A. Vulpiani, "Data science and the art of modelling", *Lettera Matematica International*. 2018.
9. Li Z, Yi Y, Luo X, Xiong N, Liu Y, Li S, et al. "Development and Clinical Application of A Rapid IgM-IgG Combined Antibody Test for SARS-CoV-2 Infection Diagnosis". *Journal of medical virology*. 2020
10. Ch. Meyer, W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press. 1995.
11. J. Pomerantz. *Metadata*. MIT Press Ltd. 2015.
12. B. Shanmugaraj, A. Malla, W. Phoolcharoen. "Emergence of Novel Coronavirus 2019- nCoV: Need for Rapid Vaccine and Biologics Development". *Pathogens*. 9(2):148. 2020
13. D. Tkaczyk, P. Szostek, P. Jan Dendek, M. Fedoryszak, Ł. Bolikowski

14. "CERMINE — automatic extraction of metadata and references from scientific literature". *Conference: 2014 11th IAPR International Workshop on Document Analysis Systems*. 2014.
15. D. Tkaczyk, P. Szostek, P. Jan Dendek, M. Fedoryszak, Ł. Bolikowski
16. "CERMINE — automatic extraction of metadata and references from scientific literature", *International Journal on Document Analysis and Recognition (IJDAR)*. Springer. 2015
17. H. van Ditmarsch, W. van der Hoek, B. Kooi. *Dynamic Epistemic Logic*, Synthese Library, Volume 337. Netherlands: Springer. 2007.
18. H. van Ditmarsch, J. Halpern, B. Van Der Hoek, B. Kooi. *Handbook of Epistemic Logic*. College Publications. 2015.