Science & Philosophy Volume 9(2), 2021, pp. 83-96

Epistemic logic for metadata modelling from scientific papers on COVID-19

Simone Cuconato♦

Abstract

The field of epistemic logic developed into an interdisciplinary area focused on explicating epistemic issues in, for example, artificial intelligence, computer security, game theory, economics, multiagent systems and the social sciences. Inspired, in part, by issues in these different 'application' areas, in this paper I propose an epistemic logic **T** for metadata extracted from scientific papers on COVID-19. More in details, I introduce a structure $\mathcal{S}$ to syntactically and semantically modelling metadata extracted with systems for extracting structured metadata from scientific articles in a born-digital form. These systems will be considered, in the logical model created, as 'Metadata extraction agents' (MEA). In this case MEA taken into consideration are CERMINE and TeamBeam. In an increasingly data-driven world, modelling data or metadata means to help systematise existing information and support the research community in building solutions to the COVID-19 pandemic.

Keywords: epistemic logic; applied non-classical logics; metadata modelling; COVID-19 pandemic

♦ Department of Computer Science, Modelling, Electronics and Systems Engineering, University of Calabria, Via P.Bucci, 87036, Rende (CS), Italy. simone.cuconato@unical.it

† Received: 2021-09-01; Accepted: 2021-12-25; Published: 2021-12-31; doi: 10.23756/sp.v9i2.652. ISSN 2282-7757; eISSN 2282-7765.

.

1. Role of (meta)data in managing COVID-19 pandemic

We are living in the age of big data, advanced analytics, and data science. The art of data science [5] has attracted increasing interest from a wide range of domains and disciplines. In the last few decades, the advent of computers and later the World Wide Web (WWW) has changed human civilization in a radical way. Now we live in a world which is being overloaded with data and information. WWW has also influenced the overall growth in scientific literature. In the light of a report issued by International Association of Scientific, Technical and Medical Publishers, there is an increase in publishing scientists by 4–5% annually. Naturally, the situation intensified during the pandemic. In just twelve months, major databases have been flooded with research articles, letters, reviews, notes, and editorials related to COVID-19[1].

As the number of scientific literature increases quickly, getting access to the core information of scientific papers easily and fast is becoming more and more important. With this core information, we can improve both the quality and efficiency of information retrieval, literature search engine and research trend prediction. In the information world, at the most elementary level, metadata are defined as 'data about data' [12, 15]. Metadata is broadly classified into three types by NISO (2004) [13] that includes descriptive, structural and administrative. Descriptive metadata is used for discovery and identification, structural metadata helps in determining how a paper is organized, while administrative metadata provides information regarding resource management. In the context of research articles, metadata is usually of descriptive nature. It provides a brief overview of a scientific article by providing information such as the title of an article, its authors and keywords etc. Hence, researchers tend to decide paper relevance with their domain of interest-based on metadata information such as title, abstract, references, authors, citing articles and affiliations. In addition to that, digital research repositories also make use of metadata in order to provide support regarding literature acquisition for the research community. Ultimately, whether descriptive, administrative or structural, metadata share a single multifunctional goal: to contribute to a clearer and more modular management

[1] It is estimated that 23,634 unique published articles have been indexed on Web of Science and Scopus between 1 January and 30 June 2020.

of digital objects and content retrieval. Automated metadata extraction enables the direct extraction of metadata from document sources. However, metadata extraction is a complicated task and poses the following challenges:

- It is hard to determine whether an extracted item from a scientific paper is representative or not.
- To the best of our knowledge, there is not a public labeled dataset, even an effective and widely accepted annotation rules.
- Although all scientific papers follow a common writing rule, the metadata may be flexible enough to appear in any section, making the metadata extraction very challenging.

Various tools and frameworks exist to automatically extract this information from PDF documents. Systems such as CERMINE or TeamBeam, for example, are able to automatically extract metadata from specific document sources.

CERMINE [16, 17] is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form. The system is based on a modular workflow and the implementations of most steps are based on supervised and unsupervised machine-learning techniques.

The TeamBeam algorithm [10] has been developed to provide a flexible tool to extract a wide array of meta-data from scientific articles. At its core, TeamBeam is a supervised machine learning algorithm, where labelled training examples are used to learn a classification scheme for the individual text elements of an article. The main goal of this paper is the modelling of metadata extracted from scientific papers on COVID-19 through the application of epistemic logic [3].

2. Epistemic Logic: Syntax, Semantics and Axioms

Since Hintikka's [6] epistemic logic [18, 19], the logic of knowledge, has been a subject of research in philosophy [7], computer science [4], artificial intelligence [11] and game theory[1, 9]. 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:

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

Definition 2.1 [Syntax of $\mathcal{L}_K$] The epistemic language $\mathcal{L}_K$ is defined as follows:

$$\varphi := p|\neg\varphi|\varphi \wedge \varphi|K_a\varphi$$

where $p \in \mathcal{P}$, $a \in \mathcal{A}$, $\mathcal{A}$ is a finite set of agents, and $\mathcal{P}$ is a countable set of atomic sentences.

Besides the standard Boolean operators, this language contains the epistemic constructions $K_a\varphi$ which we read as 'agent a knows (that) φ'. Note that an agent may be a human being, a player in a game, a robot, a machine, a 'process', or in our case a 'Metadata extraction agent' (MEA).

To build an interpretation, I first introduce the concept of an epistemic model, given by a set of possible worlds and, for each agent a in a given finite set $\mathcal{A}$, a binary relation, representing agent a's subjective epistemic indistinguishability:

Definition 2.2 [Epistemic Model] Given a set $\mathcal{P}$ of primitive propositions and a set $\mathcal{A}$ of agents, an epistemic model is a structure $M: \langle W, R^{\mathcal{A}}, V^{\mathcal{P}} \rangle$ where

- $W \neq \emptyset$ is a set of possible worlds;
- $R^{\mathcal{A}}$ is a function, yielding an accessibility relation $R_a \subseteq W \times W$ for each agent $a \in \mathcal{A}$;
- $V^{\mathcal{P}}: W \to (\mathcal{P} \to \{true, false\})$ is a function that, for all $p \in \mathcal{P}$ and $w_i \in W$, determines what the truth value $V^{\mathcal{P}}(w_i)(p)$ of p is in world w.

Definition 2.3 [Semantics of $\mathcal{L}_K$]: Given a model $M: \langle W, R^{\mathcal{A}}, V^{\mathcal{P}} \rangle$, I define what it means for a formula φ to be true in (M, w_i), written $M, w_i \vDash \varphi$, inductively as follows:

$M, w_1 \vDash p$	iff	$V(w_1)(p) = true$ for $p \in \mathcal{P}$
$M, w_1 \vDash \varphi \wedge \psi$	iff	$M, w_1 \vDash \varphi$ and $M, w_1 \vDash \psi$
$M, w_1 \vDash \neg\varphi$	iff	not $M, w_1 \vDash \varphi$ (often written $M, w_1 \nvDash \varphi$)
$M, w_1 \vDash K_a\varphi$	iff	$M, w_2 \vDash \varphi$ for all w_2 such that $w_1 R_a w_2$

Definition 2.4 [Axioms and Inference Rules] The proof system of epistemic logic that I use is axiomatized by using the axiom of T and the rule of modus ponens and necessitation. The full system is presented in Table 1:

K	$\vdash K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$
T	$\vdash K_a\varphi \to \varphi$
MP	if $\vdash \varphi \to \psi$ and $\vdash \varphi$, then ψ
NEC	if $\vdash \varphi$, then $K_a\varphi$

Table 1

The reflexivity of R guarantees that the principle

T $K_a\varphi \to \varphi$

is valid.

3. The 'Metadata Extraction Logic' Model

Let us now see how to adapt the standard epistemic logic to metadata modelling. At the syntactic level I use only one particular kind of proposition $p_{\mathcal{E}}$

$$p_{\mathcal{E}} =_{def} \mathcal{E}_{m_i}^{d_i}$$

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

Definition 3.1 [Syntax of $\mathcal{L}_{K_{\mathcal{E}}}$] Let $\mathcal{P}_{\mathcal{E}}$ be a set of primitive propositions and $\mathcal{F}$ a set of framework symbols. Then I define the language $\mathcal{L}_{K_{\mathcal{E}}}$ by the following BNF:

$$\varphi \ ::= p_{\mathcal{E}} | \neg\varphi | \varphi \wedge \varphi | K_a\varphi$$

where $p_{\mathcal{E}} \in \mathcal{P}_{\mathcal{E}}$ and $a \in \mathcal{F}$.

On a semantic level I replace the concept of possible world with that of *possible extraction*.

Definition 3.2 [Epistemic Model] Given a set $\mathcal{P}_{\mathcal{E}}$ of primitive propositions and a set $\mathcal{F}$ of frameworks/MEA, an epistemic model is a structure $M: \langle E, R^{\mathcal{F}}, V^{\mathcal{P}_{\mathcal{E}}} \rangle$ where

- $E \neq \emptyset$ is a set of possible extractions;
- $R^{\mathcal{F}}$ is a function, yielding an accessibility relation $R_a \subseteq E \times E$ for each agent

$a \in \mathcal{F}$;

- $V^{\mathcal{P}_\mathcal{E}}: E \to (\mathcal{P}_\mathcal{E} \to \{true, false\})$ is a function that, for all $p_\mathcal{E} \in \mathcal{P}_\mathcal{E}$ and $e_i \in E$, determines what the truth value $V^{\mathcal{P}_\mathcal{E}}(e_i)(p_\mathcal{E})$ of $p_\mathcal{E}$ is in extraction e.

Definition 3.3 [Semantics of $\mathcal{L}_{K_\mathcal{E}}$]: Given a model $M: \langle E, R^\mathcal{F}, V^{\mathcal{P}_\mathcal{E}} \rangle$, I define what it means for a formula φ to be true in (M, e_i), written $M, e_i \vDash \varphi$, inductively as follows:

$M, e_1 \vDash p_\mathcal{E}$	iff	$V(e_1)(p_\mathcal{E}) = true$ for $p_\mathcal{E} \in \mathcal{P}_\mathcal{E}$
$M, e_1 \vDash \varphi \wedge \psi$	iff	$M, e_1 \vDash \varphi$ and $M, e_1 \vDash \psi$
$M, e_1 \vDash \neg\varphi$	iff	not $M, e_1 \vDash \varphi$
$M, e_1 \vDash K_a\varphi$	iff	$M, e_2 \vDash \varphi$ for all e_2 such that $e_1 R_a e_2$

Definition 3.4 [Epistemic Metadata Extraction Structure] A $\mathcal{S}$ structure is of the form $\mathcal{S} = \langle \mathcal{F}, E, \mathcal{P}_\mathcal{E}, M, D \rangle$, where:

$\mathcal{F} = \{a, b, c, \dots\}$ is a non-empty finite set of MEA,

$E = \{e_1, \dots, e_m\}$ is a non-empty set of possible extractions ($|E| = m \in \mathbb{N}$),

$\mathcal{P}_\mathcal{E} = \{p_{\mathcal{E}_1}, \dots, p_{\mathcal{E}_m}\}$ is a non-empty set of propositions ($|\mathcal{P}_\mathcal{E}| = m \in \mathbb{N}$),

$M = \{m_1, \dots, m_m\}$ 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 structure in which possible extractions E occur. $\mathcal{F}$ is the set of MEA, while $\mathcal{P}_\mathcal{E}$ is the set of epistemic propositions. *M* is the set of metadata and *D* is the set of documents (papers on COVID-19).

I define, in more detail, how it is possible to systematically determine the truth value of a formula in the structure $\mathcal{S}$. In propositional logic, whether p is true or not 'depends on the situation'. In $\mathcal{S}$ a proposition $p_\mathcal{E}$ 'is true in e on condition that it is true in all possible extractions accessible from e', and since $p_\mathcal{E}$ has the form $\mathcal{E}_{m_i}^{d_i}$ I write that it is true (T) or false (F) that 'in the extraction e_i a MEA extracts the metadata m_i from the document d_i' as follows

$$\underbrace{\mathcal{E}_{m_i}^{d_i}}_{e_i} = \text{T/F}$$

Definition 3.5 [Axioms and Inference Rules] The proof system of metadata extraction logic model that I use is axiomatized by using the axiom of T and

the rule of modus ponens and necessitation. The system is presented in Table 2:

System	Rules	Axioms	Relation *R*	Figure
T	MP and NEC	$K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$ $K_a\varphi \to \varphi$	R is reflexive	$\overset{a}{\curvearrowright}$ $\underbrace{\mathcal{E}_{m_i}^{d_i}}_{e_i}$

Table 2

For example, in the graph we have a situation in which given an input document and two metadata, a MEA knows that four possible extractions can occur: the extraction in which both metadata are correctly extracted, the extraction in which metadata one is correctly extracted while metadata two is not, the extraction in which metadata two is correctly extracted while metadata one is not, and finally the extraction in which both metadata are not correctly reported.

$$\overset{a}{\curvearrowright} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}}_{e_1} \qquad \overset{a}{\curvearrowright} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}}_{e_2}$$

$$\overset{a}{\curvearrowright} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}}_{e_3} \qquad \overset{a}{\curvearrowright} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}}_{e_4}$$

4. Metadata Modelling

Let us now consider that we need to extract four metadata – title, author, keywords and journal – from three documents/scientific articles using two different MEA: CERMINE framework a and TeamBeam algorithm b.The first document d_1 concerns a medical article presenting the progress of scientific knowledge in the first five months after the start of the pandemic[8]. The second document d_2 concerns Italian research focused on the development of 'monoclonal-type' plastic antibodies based on Molecularly Imprinted Polymers (MIPs) able to selectively bind a portion of the novel coronavirus SARS-CoV-2 spike protein to block its function and, thus, the infection

process [14]. The latest document d_3 concerns a comprehensive quantitative analysis of Omicron's infectivity, vaccine-breakthrough, and antibody resistance [2].

Consider the following structure $\mathcal{S} = \langle \mathcal{F}, E, \mathcal{P}_{\mathcal{E}}, M, D \rangle$:

$\mathcal{F} = \{a, b\}$;

$E = \{e_1, \dots, e_m\}$;

$\mathcal{P}_{\mathcal{E}} = \{p_{\mathcal{E}_1}, \dots, p_{\mathcal{E}_m}\}$

$M = \{m_1, m_2, m_3, m_4\}$

$D = \{d_1, d_2, d_3\}$

Given document $d_1$and MEA a the following scenario occurs:

$$\overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_1} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_2} \qquad \overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_3} \qquad \overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_4}$$

$$\overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_5} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_6} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_7} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_8}$$

$$\overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_9} \qquad \overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_{10}} \qquad \overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{11}} \qquad \overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{12}}$$

$$\overset{a}{\rightsquigarrow} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{13}} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{14}} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{15}} \qquad \overset{a}{\rightsquigarrow} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{16}}$$

MEA a correctly extracts all metadata and therefore extraction e_1 occurs:

- $\underbrace{\mathcal{E}_{m_1}^{d_1}}_{e_1} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_2}^{d_1}}_{e_1} = \mathrm{T}$

- $\underbrace{\mathcal{E}_{m_3}^{d_1}}_{e_1} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_4}^{d_1}}_{e_1} = \mathrm{T}$

Given document d_1 and MEA b the following scenario occurs:

$$\overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_1} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_2} \qquad \overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_3} \qquad \overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_4}$$

$$\overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_5} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_6} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_7} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_8}$$

$$\overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_9} \qquad \overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \mathcal{E}_{m_4}^{d_1}}_{e_{10}} \qquad \overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{11}} \qquad \overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{12}}$$

$$\overset{b}{\sim} \underbrace{\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{13}} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{14}} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{15}} \qquad \overset{b}{\sim} \underbrace{\neg\mathcal{E}_{m_1}^{d_1}, \neg\mathcal{E}_{m_2}^{d_1}, \neg\mathcal{E}_{m_3}^{d_1}, \neg\mathcal{E}_{m_4}^{d_1}}_{e_{16}}$$

MEA b does not correctly extract the metadata m_3 and therefore extraction e_3 occurs:

- $\underbrace{\mathcal{E}_{m_1}^{d_1}}_{e_3} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_2}^{d_1}}_{e_3} = \mathrm{F}$
- $\underbrace{\mathcal{E}_{m_3}^{d_1}}_{e_3} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_4}^{d_1}}_{e_3} = \mathrm{T}$

With the first document d_1, MEA a correctly extracts all metadata, while MEA b extracts three out of four metadata. These extractions can be represented by the model

of Figure 1

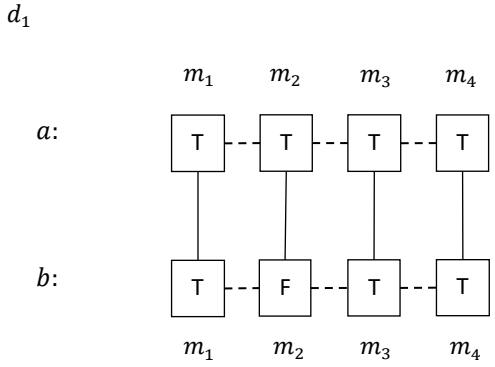

Figure 1: The model of $\mathcal{S}$ in d_1

With the second document d_2 using MEA a the extraction e_4 is realised, while with MEA b the extraction e_2 is realised

a:	b:
$\underbrace{\mathcal{E}_{m_1}^{d_2}}_{e_4} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_1}^{d_2}}_{e_2} = \mathrm{F}$
$\underbrace{\mathcal{E}_{m_2}^{d_2}}_{e_4} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_2}^{d_2}}_{e_2} = \mathrm{T}$
$\underbrace{\mathcal{E}_{m_3}^{d_2}}_{e_4} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_3}^{d_2}}_{e_2} = \mathrm{T}$
$\underbrace{\mathcal{E}_{m_4}^{d_2}}_{e_4} = \mathrm{F}$	$\underbrace{\mathcal{E}_{m_4}^{d_2}}_{e_2} = \mathrm{T}$

These metadata extractions can be represented by the model of Figure 2

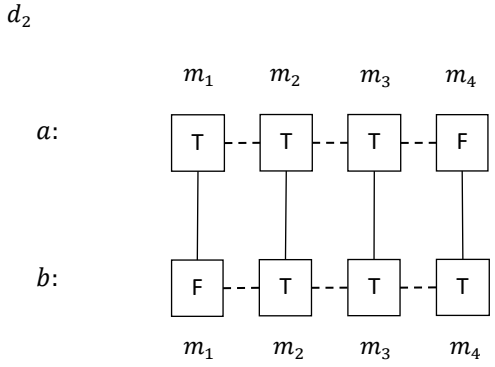

Figure 2: The model of $\mathcal{S}$ in d_2

Lastly, with the third document d_3 using MEA a the extraction e_1 is realised, while with MEA b the extraction e_{16} is realised

a:	b:
$\underbrace{\mathcal{E}_{m_1}^{d_2}}_{e_1} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_1}^{d_2}}_{e_{16}} = \mathrm{F}$
$\underbrace{\mathcal{E}_{m_2}^{d_2}}_{e_1} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_2}^{d_2}}_{e_{16}} = \mathrm{F}$
$\underbrace{\mathcal{E}_{m_3}^{d_2}}_{e_1} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_3}^{d_2}}_{e_{16}} = \mathrm{F}$
$\underbrace{\mathcal{E}_{m_4}^{d_2}}_{e_1} = \mathrm{T}$	$\underbrace{\mathcal{E}_{m_4}^{d_2}}_{e_{16}} = \mathrm{F}$

These metadata extractions can be represented by the model of Figure 3

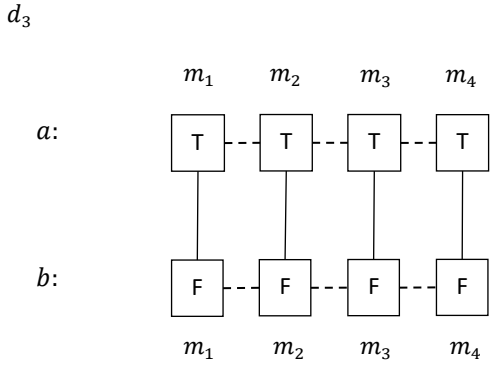

Figure 3: The model of $\mathcal{S}$ in d_3

The models seen above were focused on the representation of a single document. However, a fundamental aspect of metadata modelling is to be able to focus on the single metadata. For this reason, I propose a second representation of the extracted metadata in $\mathcal{S}$. In Figure 4, 5, 6 and 7 I highlight how the extraction systems behaved in each single extraction. When the metadata is reported correctly the box is white, while when it is reported incorrectly then the box is grey.

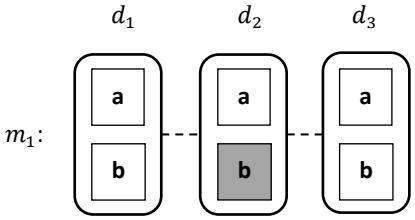

Figure 4

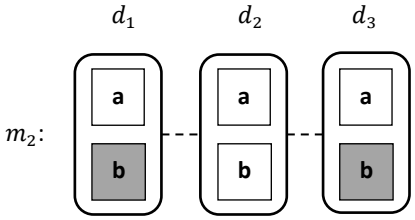

Figure 5

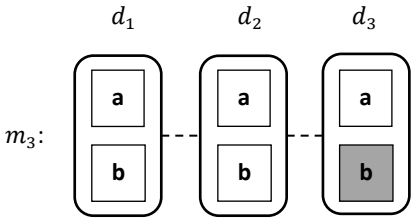

Figure 6

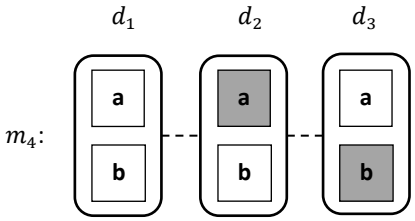

Figure 7

5. Conclusions

There is no doubt that the potential of data science and analytics to enable data-driven theory, economy, and professional development is increasingly being recognized. This involves not only core disciplines such as computing, informatics, and statistics, but also logic, ethic or the broad-based fields of business, social science, and health/medical science. However, one should be mindful that data without a model is just noise. Motivated by the preceding concerns and observations, in this paper I have presented a logical modelling of metadata extracted from scientific papers on COVID-19. In an increasingly data-driven world, modelling data or metadata means to help systematise existing information and support the research community in building solutions to the COVID-19 pandemic.

References

[1] Bonanno, P., Battigalli, G. 1999, 'Recent results on belief, knowledge and the epistemic foundations of game theory', *Research in Economics*, Volume 53, Issue 2, pp. 149-225.

[2] Chen, J., Wang, R., Benovich Gilby N., Wei G-W., 'Omicron (B.1.1.529): Infectivity, vaccine breakthrough, and antibody resistance' (*forthcoming*).

[3] Cuconato, S. 2021, 'Epistemic logic and CERMINE: a logical model for automatic extraction of structured metadata', *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 9 (1), pp. 161-172.

[4] Fagin, R., Halpern, J. Y., Moses, Y., Vardi, M. Y. 1995, *Reasoning About Knowledge*. The MIT Press: Cambridge.

[5] Graham, M. J. 2012, 'The art of data science', In *Astrostatistics and Data Mining*, Springer Series in Astrostatistics, Vol. 2. 47—59.

[6] Hintikka, J. 1962, *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.

[7] Hintikka, J. 1986, 'Reasoning about knowledge in philosophy', In J. Y. Halpern, editor, *Procedings of TARK*, Morgan Kaufmann: San Mateo, CA.

[8] Hua Li, Zhe Liu, Junbo Ge, 2020, 'Scientific research progress of COVID-19/SARS-CoV-2 in the first five months', *Journal of Cellular and Molecular Medicine*.

[9] Jackson M. and Zenou, Y. 2014, 'Games on networks', In Peyton Young and Shmuel Zamir, editors, *Handbook of Game Theory*, volume 4. Elsevier Science.

[10] Kern, R., Jack, K., Hristakeva, Granitzer, M. 2012, 'TeamBeam - Meta-Data Extraction from Scientific Literature', *Computer Science D Lib Mag*.

[11] Meyer, Ch. and van der Hoek, W. 1995, *Epistemic Logic for AI and Computer Science*, Cambridge University Press.

[12] J. Pomerantz. *Metadata*, MIT Press Ltd. 2015.

[13] NISO 2004, *Understanding metadata*, 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814 USA: NISO

[14] Puoci, F., Parisi, O. I. 2020, ''Monoclonal-type'' plastic antibodies for SARS-CoV-2 based on Molecularly Imprinted Polymers' (*forthcoming*).

[15] Rovella, A. 2019, 'Metadata consistency and coherence in the digital management and preservation process of administrative records', *AIDAInformazioni*, numero 1-2, pp. 75-97.

[16] Tkaczyk, D., Szostek, P., Jan Dendek, P., Fedoryszak, M., Bolikowski. Ł. 2014, 'CERMINE — automatic extraction of metadata and references from scientific literature', *Conference: 2014 11th IAPR International Workshop on Document Analysis Systems*.

[17] Tkaczyk, D., Szostek, P., Jan Dendek, P., Fedoryszak, M., Bolikowski. Ł. 2015, 'CERMINE — automatic extraction of metadata and references from scientific literature', *International Journal on Document Analysis and Recognition (IJDAR)*. Springer.

[18] van Ditmarsch, H., van der Hoek, W. and Kooi, B. 2007, *Dynamic Epistemic Logic*, Synthese Library, Volume 337, Netherlands: Springer.

[19] van Ditmarsch, H., Halpern, J., van Der Hoek, W. and Kooi, B. 2015, *Handbook of Epistemic Logic*, College Publications.