UNIVERSITÀ DELLA CALABRIA

UNIVERSITÀ DELLA CALABRIA

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Dottorato di Ricerca in

Information and Communication Technologies
(*curriculum*: Computer Science and Engineering)

Ente finanziatore

Istituto di Informatica e Telematica (IIT) - CNR

CICLO

XXXV

A LOGICAL AND ONTOLOGICAL FRAMEWORK FOR METADATA EXTRACTION AND MODELLING FROM HETEROGENEOUS DOCUMENT SOURCES

Settore Scientifico Disciplinare M-STO/08

Co-Settore Scientifico Disciplinare MAT/01

Coordinatore: Ch.mo Prof. Giancarlo Fortino

Firma ____ FORTINO GIANCARLO 30.06.2023 16:51:28 UTC ____

Supervisore/Tutor: Prof.ssa Antonietta Folino

Firma____ Antonietta Folino 29.06.2023 11:44:37 GMT+01:00 ____

Supervisore/Tutor: Dott.ssa Elena Cardillo

Firma_ elena cardillo 29.06.2023 13:08:36 GMT+00:00 ____

Dottorando: Dott. Simone Cuconato

Firma ____ Simone Cuconato 30.06.2023 19:37:53 GMT+01:00 ____

Ai miei genitori, Giovanni Cuconato e Brunella Fanello

Logic is not a body of doctrine,
but a mirror-image of the world.
Logic is transcendental.

Ludwig Wittgenstein
Tractatus logico-philosophicus

Contents

List of Logic Symbols

Symbol	Logic Name	Read as	Category
¬	negation	not	propositional logic
∧	conjunction	and	propositional logic
∨	(inclusive) disjunction	or	propositional logic
⊻	(exclusive) disjunction	xor	propositional logic
→	material implication	if ... then	propositional logic
↔	material equivalence	if and only if (iff)	propositional logic

∀	universal quantification	for all,
		for any
			first-order logic

∃	existential quantification	there exists	first-order logic
∃!	uniqueness quantification	there exists exactly one	first-order logic
∄	negated quantification	There does not exist	first-order logic
⊤	tautology	tautology, truth	propositional logic, first-order logic
⊥	contradiction	contradiction, falsity	propositional logic, first-order logic
:=	definition	is defined as	everywhere
⊢	syntactic turnstile	proves	propositional logic, first-order logic
⊬	negated syntactic	not proves	propositional logic,

	turnstile		first-order logic
$\vDash$	semantic turnstile	models	propositional logic, first-order logic
$\nvDash$	negated semantic turnstile	not models	propositional logic, first-order logic
$\Diamond$	diamond operator	it is possible that	modal logic
$\Box$	box operator	it is necessary that	modal logic
K	epistemic operator	knows (that)	epistemic logic
$\mathcal{E}^{d_i}_{m_i}$	extraction predicate	extracts metadata m_i from document d_i	epistemic logic
$\sim$	classical negation	not	four-valued epistemic logic

Introduction[♦]

Together with the disruptive development of modern sub-symbolic approaches to artificial intelligence (AI), such as machine learning (ML), symbolic approaches to classical AI, based on the formal representation of knowledge and its elaboration via explicit reasoning rules, are re-gaining momentum, as more and more researchers exploit their potential to make AI more comprehensible, explainable, and therefore trustworthy. Accordingly, logic-based technologies have played over the years and are going to play a key role in the forthcoming AI landscape – in particular, for the knowledge engineering and library and information science. Along this line, the purpose of this dissertation is to build a logical and ontological framework for metadata extraction and modelling from heterogeneous document sources.

The word “metadata” is a deliberate play on Aristotle’s *Metaphysics*. But Aristotle himself did not use that title or even describe his field of study as “metaphysics”. The title “metaphysics”- literally, “after the *Physics*”- subsequent to the arrangement of Aristotle's works by Andronicus of Rhodes in the first century BC. Similarly, the word “metadata” indicates something that is beyond the data. More specifically, metadata is a means of representing the complexity of an object in a simpler

[♦] Ph.D. funded by CNR, CNR-IIT-Cosenza.

form. Metadata play a fundamental role in data science, as they provide a *criterion of identity* for data. Inspired in Willard Van Orman Quine's well-known slogan “No entity without identity”, it is possible to say “No data without metadata”.

Over the years, several metadata extraction systems have been developed. However, most metadata extraction systems generally only work with a specific document source. For this reason, the first question asked in the dissertation is: *How can we build a framework capable of extracting metadata from different document sources?* The answer is as follows: since the framework will have to be able to manage different sources and different formats, it will necessarily have to *make decisions* on the basis of the possible different document sources. For example, if the input document is a text in PDF format then the decision rules will lead to certain systems, tools and metadata, on the contrary, if the input is an image, an audio or a video, then the choices regarding systems, tools and metadata will be different. Therefore, it was built a framework, by the name MADME (MAke Decision for Metadata Extraction), as a decision system capable of performing reasoning and making decisions grounded on a *decision-making ontology* (DMO), based on first-order logic.

The second question asked in the dissertation is: *How formal models can be created of the extracted metadata?* This question was answered by proposing a model based on non-classical logic, and in particular on epistemic logic, capable of formal representation of the extracted metadata. The model was developed by introducing a new and specific predicate $\mathcal{E}$ – reads “extract” – and a structure $\mathcal{S}$ to syntactically and semantically define metadata extracted with any automatic metadata extraction system. These systems are considered, in the logical model created, as metadata extraction agents (MEA).

In this way, the *theoretical* nature of the constructed framework and the importance of the dialogue between different scientific fields such as logic, knowledge engineering, applied ontology and library and information science becomes evident.

The dissertation is organized as follows. Chapter 1 introduces the world of metadata from the general notion of theory of knowledge to the more specific notion of knowledge organization. Subsequently, the main features of metadata and metadata extraction systems are presented.

Chapter 2 aims to provide the logical tools and the conceptual issues indispensable for the construction of the proposed logical-ontological framework, also fixing the formal notation for the dissertation. In the first part, after the presentation of some useful examples of applications of classical and non-classical logic to artificial intelligence, computer science and knowledge engineering, the focus will be provide a concise introduction of classical logic. In detail, after the description of the syntax and semantics of the propositional calculus, a simplified and optimised version of semantic tableaux as a *decision procedure* will be presented. In the second part, non-classical logic will be introduced. In particular, the basics of modal logic will be described, indispensable for a correct understanding of epistemic logic, which will be used in the proposed approach to metadata modelling.

Chapter 3 aims to describe the implementation of MADME (MAke Decision for Metadata Extraction). The proposed logical-ontological approach based on three elements: decision-making (DM) ontology (DMO), DM rules (DMR) and DM procedure (DMP). DMO provides an informal and formal representation of digital objects. DMR: *i*) are derived from DMO, and *ii*) are formal rules written in the language of first-order logic (FOL) that define all decision steps in detail. The DMP provides a set of methodological guidelines for the application of DMR. MADME can be defined as a heavyweight ontology, since it includes classes, subclasses, relazionships between classes, istances, axioms, constraints, theorems and, especially, decision rules. The main objective of the MADME approach is to develop a formal decision-making ontology that can guide the choice of metadata extraction systems. In the last part of the chapter, specific examples of the MADME way of operating will be described.

Finally, chapter 4 uses epistemic logic to model structured metadata, and the tools of metaontology to propose a definition of *veracity* as *truthmaker*. In particular, two different types of epistemic logic will be presented and applied: epistemic logic **T** and a *four-valued* epistemic logic (FVEL). Epistemic logic was specifically adapted to the modelling of metadata through the introduction of a new predicate $\mathcal{E}$ - read “extract” - and a structure $\mathcal{S}$ to analyze the extracted metadata syntactically and semantically.

Origin of the material

The chapters of this dissertation have either been published as articles or are currently under review. The sources of the chapters are listed below.

- Chapter 4 is based on the following articles:

  Cuconato, S. (2022), A logical-metaontological approach to the problem of (meta)data veracity in systems for automatic extraction of metadata from scientific-legal articles, *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 10:2, pp. 168–187.

  Cuconato, S. (2021b), Epistemic logic for metadata modelling from scientific papers on COVID-19, *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 9:2, pp. 83–96.

  Cuconato, S. (2021a), 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.

- Chapter 4 is based on the following paper:

  Cuconato, S. (2023), A Four-Valued Epistemic Logic for Metadata Modelling from Medical Articles on Pain Therapies, *5th International Conference on Computational Intelligence in Pattern Recognition, Special Session: Intelligent Approaches for Data Mining Applications* (forthcoming).

- Chapter 4 is also partially based on the following articles:

  Cuconato, S. (2023), Oggetti matematici non-esistenti come truthmakers: Meinonghianismo strong e l’argomento di indispensabilità, *Rivista di Estetica* (forthcoming).

Cuconato, S. (2014a), Mondi di Wittgenstein. Metaontologia del ‘Tractatus’ e teoria dei ‘truthmakers’ di Armstrong, *Rivista Italiana di Filosofia Analitica junior*, 5:2, Special Issue: Metaphysics, pp. 53–65.

- Chapter 3 is partially based on the following monograph and article:

Cuconato, S. (2022a), *Impegno ontologico e l’argomento di indispensabilità in filosofia della matematica. Un’analisi*. Il Sileno Edizioni.

Cuconato, S. (in review), Can Nonexistent Mathematical Objects Make a Difference? Meinongianism indispensability argument and mathematical entanglement.

Chapter 1

A World of Metadata

1.1. From Theory of Knowledge to Knowledge Organization

Any fruitful discussion of knowledge does well to begin by recognizing some basic linguistic facts about how the verb *to know* and its cognates actually function in the usual range of relevant discourse. First, it is necessary to recognize that 'to know' has both a propositional and a procedural sense: there is the intellectual question of “knowing that something or other is the case” (*that*-knowledge) and the practical question of knowing how to perform an action and how to go about realizing some end (*how-to*-knowledge).

Since only the first mode of knowledge has generally been the focus of attention in traditional epistemology[1], this section will focus on specifically propositional knowledge – the kind of knowledge which is at issue in locutions to the effect that someone knows something-or-other to be the case (“x knows that p”).

The conception of “knowledge” represents a flexible and internally diverse concept. In general terms, it refers to the way in which persons can be said have access to correct information. This

[1] The term “epistemology” comes from the Greek words “episteme” and “logos”. “Episteme” can be translated as “knowledge”, while “logos” can be translated as “reason”. See Plato (2017); Williamson (2002); Rescher (2003); Steup *et al.* (2014).

can be done in rather different ways and especially depending on the kind of thing that is at issue:

- *Knowledge-that* something or other is the case (i.e., knowledge of facts). Examples: I know that 3 plus 2 is 5. I know that Rome is the capital of Italy.
- *Adverbial knowledge*. Examples: Knowing what, when, why, how, and so forth.
- *Knowledge by acquaintance* with individuals or things. Examples: I know Ramona. I know the owner of that house.
- *Performator*y (or "*how-to*") knowledge. Examples: I know how to ride a motorbike.

Traditionally epistemology, the theory of knowledge, has focused on knowledge of the first type: propositional or factual knowledge[2]. But what is propositional knowledge? In sum, propositional knowledge is a cognitive affair, and it is this aspect of knowledge that will be central to this dissertation.

The fundamental characteristics of propositional knowledge are inherent in the *modus operandi* of knowledge discourse. In particular, the following four characteristics are salient in this regard[3]:

- *Truth Commitment*. Only the truth can be known. If someone knows that *p* then *p* must be true;
- *Grounding*. Knowledge must be appropriately grounded. A person *can* accept something without a reason but cannot then be said to *know* it;

[2] Moser (1987).

[3] Cf. Rescher (2003, xvi).

- *Reflexivity*. To attribute a specific item of propositional knowledge to someone else *is ipso facto* to claim it for oneself;
- *Coherence*. Since all items of propositional knowledge must be true, they must in consequence be collectively coherent.

What is interesting and important to emphasise is that, in general, knowing requires a subject, that the cognitive content be expressed through syntactically and semantically "correct" propositions, that knowledge follows a justificatory path or method. In this way, it is possible to generate knowledge and, consequently, it is necessary to organize the acquired knowledge[4].

Knowledge organization (KO) is about describing, representing, storing and organizing documents and document representations as well as subjects and concepts both by humans and by computer programs[5]. KO is a multidisciplinary field, where concepts of library and information science, computer science, philosophy, cognitive science and linguistics, among others, meet to form an extensive body of research and practice. The two principal aspects of KO are:

> (1) knowledge organization processes (KOP) and (2) → knowledge organization systems (KOS). *Knowledge organization processes* (KOP) are, for example, the processes of cataloging, subject analysis, → indexing, → tagging and → classification by humans or computers. *Knowledge organization systems* (KOS) are the selection of concepts with an indication of selected semantic

[4] In general, on the importance of epistemology in knowledge organization and in library and information science see Hiørland (2011).
[5] Cf. Hjørland (2008). For an overview on the different aspects of the history of KO see Samurin (1964); Kedrow (1975); Hiørland (2013).

> relations. Examples are classification systems, lists of subject headings, thesauri, ontologies and other systems of metadata[6].

In this context, the majority of scholars have made a clear distinction between data, information and knowledge[7]. The criteria suggested to distinguish knowledge from information and data include temporal sequence (knowledge is based on information, which in turn is based on data), the role of structure, context and interpretation (knowledge is structured, contextualized and interpreted), value (knowledge is more valuable than Information and data) and the potential of action (knowledge, unlike information, can be directly acted upon). In summary:

- Data is directly observable
- Information represents analyzed data
- Knowledge is actionable information

Information or knowledge that is organized, stored, managed or shared requires a particular type of meta-information or meta-knowledge: metadata. Metadata emphasize meta-information or meta-knowledge aspects in that they describe the content, quality, condition, and other characteristics of other data or information[8].

1.2. Metadata Characteristics

There are several definitions of metadata. One that summaries the key points of most of these definitions is the following:

> Metadata is pervasive in information systems, and comes in many forms. The core features of most software packages we use every day

[6] Hiørland (2016).

[7] For example, see Davenport, Prusak (2000, pp. 2–6); Rollett (2003, pp. 5–6); Jashapara (2004, pp. 9–11); Martin (2008, pp. 386–387).

[8] EI-Sherbml, K1im (2004, p. 239).

> are metadata-driven. People listen to music through Spotify; post photos on Instagram; locate video on YouTube; manage finances through Quicken; connect with others via email, text, and social media; and store lengthy contact lists on their mobile devices. All of this content comes with metadata—information about the item's creation, name, topic, features, and the like. Metadata is key to the functionality of the systems holding the content, enabling users to find items of interest, record essential information about them, and share that information with others[9].

Since metadata is a broad term, it includes many types of structured "data about data". The first use of the notion of metadata can be traced back to antiquity, with its appearance in the first libraries[10], while Eden[11] claims that its purpose and meaning have been around as long as humankind.

In the twentieth century cataloguing codes became more elaborate, tipically prepared by a professional committee. A milestone in cataloguing was the publication of the Anglo-American Cataloguing Rules (Second Edition (AACR2)) in 1978[12]. At that point, a very structured metadata schema was developed, and bibliographic principles were established. The first published use of the word "metadata" in the sense of "data about data" most likely dates back in the first edition of NASA's Directory Interchange Format Manual published in 1988. Since then, the term has been widely used in the sense of information needed to make computer documents easier to manage.

At the beginning of the twenty-first century there are two main approaches to metadata that have emerged from computer

[9] NISO (p. 3).

[10] Chan (1994, p. 6).

[11] Eden (2002, p. 6).

[12] Taylor (2004, p. 59).

science and library science, namely bibliographic control and data management, respectively[13]. The data management approach deals mainly with the technical aspects of metadata, such as data security, data sharing and data integrity. The bibliographic control approach focuses on the development of information systems to organize and provide access to large collections of objects containing information.

The main purpose of metadata is to serve as a tool for the effective organization and management of information objects, which may include data, information or knowledge. Broadly speaking, information objects have three characteristics: content, context and structure; all these characteristics can be reflected through metadata:

- Content refers to what the object contains or concerns and is intrinsic to an information object;
- Context indicates the “who”, “what”, “why”, “where” and "how" aspects associated with the object's creation and is extrinsic to an information object;
- Structure refers to the formal set of associations between individual information objects and can be intrinsic or extrinsic[14].

There is no single international standard for metadata, but many application domain-specific standards. All of these have different characteristics and attributes. Metadata can come from two sources: internal metadata generated by the creating agent of an information object and external metadata that is created later, often by agents other than the object creator. Furthermore, there

[13] Burnett *et al*. (1999).
[14] Gilhland-Swetland (1998).

are two main methods for creating metadata. The first is the manual metadata created by humans and the second is the automatic metadata generated by software[15].

Also, the nature of metadata varies. One strategy is to use elaborated and specialized schemes, such as Machine-Readable Cataloging (MARC), which is the most used scheme in libraries worldwide. The other strategy is to create a schema, such as Dublin Core, that can be used by the author of a document to create a bibliographic record[16].

In general, metadata can be static or dynamic, long-term or short-term. Static metadata are those that persist as they have been created, because they provide unchanged characteristics of the information object, while dynamic metadata change with use or manipulation of the information object, to document all the changes made on the object.

Short-term metadata are mainly transactional in nature and are therefore important for shorter periods of time. In contrast, long-term metadata are needed to ensure that the object continues to be accessible and usable. Depending on their function, we can distinguish different types of metadata. The typologies proposed by many authors[17] are the following:

- Descriptive: Metadata describes a resource for the purposes of discovery and identification of relevant information. Characteristic examples of descriptive metadata are the title, keywords or abstract of a source. It

[15] Ivi (p. 6).
[16] https://www.dublincore.org/. Dublin Core will be illustrated in the next section.
[17] Gilhland-Swetland (1998, p. 3); Eden (2002, p. 10); Caplan (2003, pp. 3–5); Haynes (2004, p. 14); Taylor (2004, pp. 147–152).

serves the same functions in resource discovery as cataloguing does by: allowing resources to be found by relevant criteria; identifying resources; bringing similar resources together; and giving location information;

- Structural: This refers to the structure and relationships of a set of digital resources. It is important because the structure of an information object is an indicator of that object's meaning. Furthermore, it can describe relationships between resources, such as the relationship between a report and an executive summary written in a different language;
- Administrative: This provides information to help manage a resource, such as when and how it was created, its file type and who can access;
- Technical: This is related to how a system functions and how metadata behave. It may include the hardware and software documentation and security data;
- Rights management: This concerns with intellectual property rights. It may include a note stating whether the content can be used outside the borders of the organization or not;
- Preservation: This contains information needed to archive and preserve a resource, such as data refreshing and migration;
- Use: This is related to the level and type of use of Information resources. In addition to use and tracking, it may contain, for instance, information on content reuse and multiple versions of content.

1.3. Dublin Core Metadata Element Set

The large number of metadata schemes developed has led to the attempt to achieve interoperability between various metadata standards. Interoperability is the main reason for creating a standard in the first place. Therefore, standards have been proposed that enable the communication between metadata.

This section describes one metadata scheme that has achieved standard status and will be used on several occasions in this dissertation: the Dublin Core Metadata Element Set (DCMES). DCMES grew out of a workshop sponsored by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications in 1995. DCMES was created to describe web documents but is general enough to also represent the content of other resources, such as text documents, images, audio and video.

DCMES is a set of descriptive elements capable of representing any information resource accessible on the network. It comprises a basic (simple) level and a level that provides more detailed (qualified) information about resources. The Simple Dublin Core consists of 15 elements. Since it can be unspecific and, in some circumstances, even ambiguous, the model has been extended, resulting in the Qualified Dublin Core, which is an extension of Simple Dublin Core through the use of additional elements, element refinements, and encoding schemes.

DCMES was developed following two fundamental characteristics: simplicity and generality, and a set of guidelines:

- Facility: The creation and maintenance of metadata must be simple. At the same time, the set of DCMES metadata must allow effective searching of information.

- Use of a universally accepted semantics: the possibility of formulating requests based on a metadata set requires that the meaning that is associated with the different elements of the metadata set be the same for the cataloguer and the request formulator.
- Possibilities of international use: The translation of DCMES is rather simple, as it only requires the translation of each element of the model and the various element descriptions into the different languages.
- Extensibility: Although the DCMES metadata set is rather limited, it is possible to extend the set to meet the needs of particular user communities.

In addition to these guidelines, the DCMES model also conforms to a number of general principles:

- Each element is optional and can be repeated: Having all elements optional makes it possible to: *i*) easily manage interoperability with other models and *ii*) simplify the verification of syntactic correctness of DCMES records;
- Each DCMES metadata record describes a manifestation of the resource;
- The presence of qualifiers may be ignored.: As we have seen, DCMES allows qualifiers to be associated with each element record describing a resource. However, to make the use of DCMES as general as possible, it is required that any qualifier may be ignored and that the element may be used as if the qualifier did not exist.

DCMES contains the following fifteen metadata

Element name: Title	Label: Title

	Definition: A name given to the resource. Comment: Typically, Title will be a name by which the resource is formally known.
Element name: Creator	Label: Creator Definition: An entity primarily responsible for making the content of the resource Comment: Examples of Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity
Element name: Subject	Label: Subject Definition: A topic of the content of the resource Comment: Typically, Subject will be expressed as keywords, key phrases, or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme
Element name: Description	Label: Description Definition: An account of the content of the resource Comment: Examples of Description include, but are not limited to, an abstract, table of contents, reference to a graphical representation of content, or free-text account of the content
Element name: Publisher	Label: Publisher

	Definition: An entity responsible for making the resource available Comment: Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity
Element name: Contributor	Label: Contributor Definition: An entity responsible for making contributions to the content of the resource Comment: Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity
Element name: Date	Label: Date Definition: A date of an event in the lifecycle of the resource Comment: Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3CDTF] and includes (among others) dates of the form YYYY-MM-DD
Element name: Type	Label: Type Definition: The nature or genre of the content of the resource Comment: Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example,

	the DCMI Type Vocabulary [DCT]). To describe the physical or digital manifestation of the resource, use the Format element
Element name: Format	Label: Format Definition: The physical or digital manifestation of the resource Comment: Typically, Format will include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats
Element name: Identifier	Label: Resource Identifier Definition: An unambiguous reference to the resource within a given context Comment: Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI), and the International Standard Book Number (ISBN)
Element name: Source	Label: Source Definition: A reference to a resource from which the present resource is derived

	Comment: The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system.
Element name: Language	Label: Language Definition: A language of the intellectual content of the resource Comment: Recommended best practice is to use RFC 3066 [RFC3066], which, in conjunction with ISO 639 [ISO639], defines two- and three-letter primary language tags with optional subtags. Examples include “en” or “eng” for English, “akk for Akkadian, and “en-GB” for English used in the United Kingdom
Element name: Relation	Label: Relation Definition: A reference to a related resource Comment: Recommended practice is to identify the related resource by means of a URI. If this is not possible or feasible, a string conforming to a formal identification system may be provided
Element name: Coverage	Label: Coverage Definition: The spatial or temporal topic of the resource, spatial applicability of the resource, or jurisdiction under which the resource is relevant Comment: Spatial topic and spatial applicability may be a named place or a location specified by its geographic

	coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended practice is to use a controlled vocabulary such as the Getty Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods may be used in preference to numeric identifiers such as sets of coordinates or date ranges. Because coverage is so broadly defined, it is preferable to use the more specific subproperties Temporal Coverage and Spatial Coverage
Element name: Rights	Label: Rights Definition: Information about rights held in and over the resource Comment: Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource

Table 1.3.1. Dublin Core metadata

In many cases, DCMES elements are unspecific and ambiguous, so the need to provide more detailed information on certain fields was highlighted. This need was considered by the group that defined the standard and is responsible for its evolution, and qualifiers were introduced that can be associated with each Dublin Core element. Currently there are two classes of qualifiers:

- Element refinement: it provides additional information in order to make the meaning of an element more specific. The use of an element refinement for a given element does not change its semantics, but better specifies its meaning.
- Encoding scheme: it makes the interpretation of the value of an element less ambiguous and clearer.

1.4. Metadata Extraction Systems

The work on Dublin Core has led to an increasing awareness of metadata throughout different application domains, from which point an uncontrollable number of domain-specific metadata element sets were published. These are commonly referred to as metadata schemes. According to Priscilla Caplan metadata schema is a "set of metadata elements and rules for their uses that has been defined for a particular purpose[18]" Different metadata extraction systems have been developed over the years.

This section describes tools and systems capable of extracting various types of metadata and content from a specific type of documentary source concerning scientific literature. The approaches differ in the scope of extracted information, methods used, input and output formats, availability, and licenses. Typically, at the beginning of document processing a layout analysis is performed and the regions of the document are

[18] Caplan (2003).

classified using various algorithms. These fragments are usually located in the document using specific rules or machine learning.

For example, Hu[19] describes a machine learning-based approach for extracting titles from general documents and as a case study, Microsoft Word and PowerPoint document are used. After pre-processing the document in specific units, units are then transformed to features and classified. Two types of features were used: format features (font size, alignment, boldface, the presence of blank lines) and linguistic features (keywords specific for part of the document, number of words).

Cui and Chen[20] describe a system for extracting metadata from PDF documents. In this case, text extraction and page segmentation is done with the use of PDF to HTML, a third-party open-source tool.

One of the most advanced extraction systems in this context and which will be used on several occasions in the dissertation is CERMINE[21]. CERMINE[22] is a comprehensive open-source system for extracting structured metadata from scientific articles in a born-digital form[23]. 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 1.4.1.[24] and 1.4.2.[25], consists of three paths (*ii* and *iii* run in parallel): *i*) the

[19] Hu *et al*. (2005).

[20] Cui, Chen (2010).

[21] The "technical" reasons for this will be noted in the last section of Chapter 3.

[22] Tkaczyk *et al*. (2014, 2015).

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

[24] Tkaczyk *et al*. (2014).

[25] Tkaczyk *et al*. (2015).

base structure extraction path requires a pdf file as input and produces a geometric hierarchical structure in TrueViz format. 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.

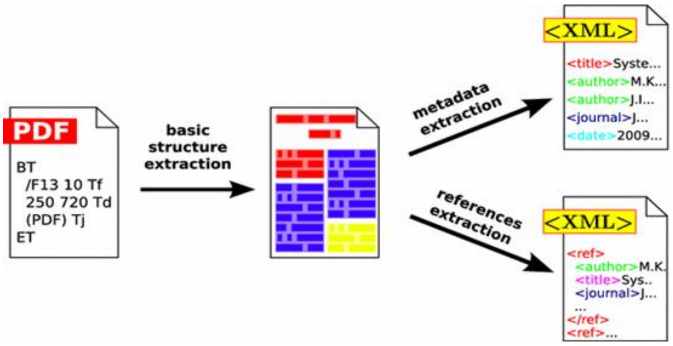
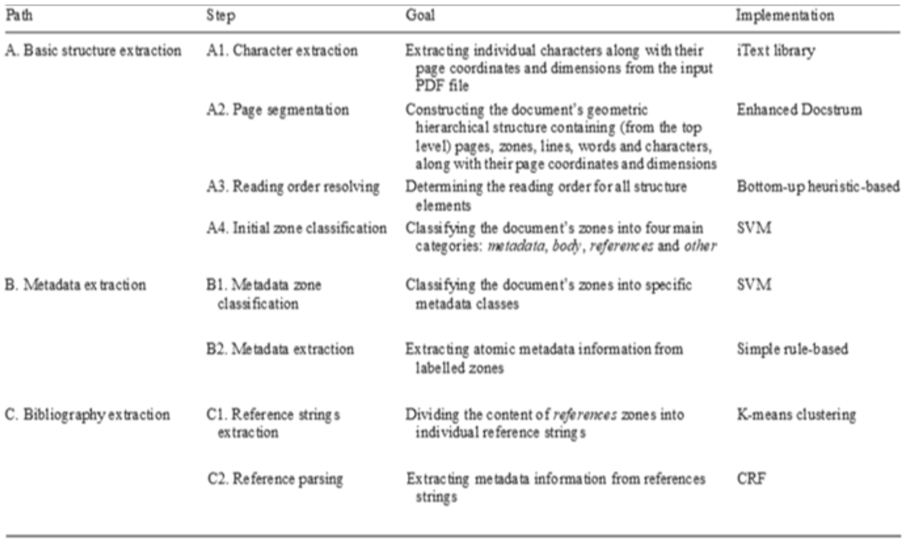


Figure 1.4.1. CERMINE's extraction workflow architecture



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 1.4.2. The decomposition of CERMINE's extraction workflow

CERMINE's core extraction algorithm makes extensive use of support vector machines (SVM). It is a powerful machine learning classification technique that can handle a wide variety of inputs and work effectively even with small training data. SVM is a linear model of the form

$$y(x) = \mathrm{w}^{T}\phi(x) + b$$

Where

- x is a feature vector representing the classification instance;
- $\phi(x)$ denotes a fixed feature-space transformation;
- w and b are parameters determined during the training based on the training instances;
- new instances are classified according to the sign of $y(x)$.

The use of SVM for classification and extraction algorithms are one of the most useful techniques for extracting information from documents. However, the use of specific algorithms and approaches to the problem of metadata extraction is generally aimed at a specific class or type of document source. On the contrary, in our dissertation the problem is addressed to heterogeneous documentary sources. Consequently, since the framework will have to be able to manage different sources and different formats, it will necessarily have to make decisions based on the possible different document sources. For this reason, our framework will not be based on classification or extraction algorithms, but on a specific decision-making procedure based on logic and ontology. For example, if the input document is a text in PDF format then the decision rules will lead to certain tools and metadata extraction systems, on the contrary if the input is an image, an audio or a video, then the choices regarding tools and metadata extraction systems will be different. Therefore, our choice was to build: *i*) a theoretical framework as a decision system capable of performing reasoning and making decisions on the basis of a decision-making ontology; and *ii*) a model based on epistemic logic to formalize structured metadata.

Chapter 2

Logical Tools for a Logical-Ontological Framework for Metadata Extraction and Modelling

2.1. Logic-Based Technologies for Intelligent Systems

The chapter aims to provide the logical tools and the conceptual issues indispensable for the construction of the proposed logical-ontological framework, also fixing the formal notation for the dissertation. Logic plays a fundamental role in computer science, and it is necessary to understand its basic concepts in order to study many of the more advanced subjects in computing. Here are just a few examples covering the whole range of computing applications:

- In artificial intelligence (AI)[26], logical languages are widely used to express the necessary declarative knowledge. Symbolic logic also provides a clear semantics for knowledge representation languages and a methodology for analyzing and comparing deductive inference techniques.
- In software engineering, it is good practice to specify what a system should do before starting to code it. Logic is often used for software specifications.

[26] Two reference books are Meyer, van der Hoek (1995); Minker (2000).

- In digital circuit design and computer architecture, logic is the language used to describe the signal values that are produced by components.
- In database systems, logic is relevant to the relational data model, where data are organized in the form of relations.
- In safety-critical applications, it is essential to establish that a program is correct. Formal logic is the foundation of program correctness proofs.
- In programming language design, one of the most commonly used methods for specifying the meaning of a computer program is the lambda calculus.
- In data science, logical modelling of data and metadata allows information to be organized and formalized.
- In computability theory, logic is used both to specify abstract machine models and to reason about their capabilities.

In particular, while it is true that *sub-symbolic* AI techniques, such as machine learning —there including deep learning and neural networks— are the most widely used techniques, it is also true that *symbolic approaches* are re-emerging in at least three respects[27]: *i*) as a means of bringing AI closer to human understanding; *ii*) as a formal study of programs and semantics in the computational model, inference as computation and automatic theorem proving; and last but not least, *iii*) as logic-based approaches in successful agent-based models and technologies: indeed, agents reason through logic, and plan and coordinate through logical processes. It is precisely in the latter sphere that the dissertation is located. In

[27] For a state of the art in logic-based technologies for intelligent systems see Calegari *et al.* (2020).

general, the logic-based technologies developed in over 70 years can be classified as technologies that meet the needs of: *i*) knowledge representation; *ii*) reasoning; and *iii*) model checking and verification. Figure 2.1.1. summarizes this classification[28]:

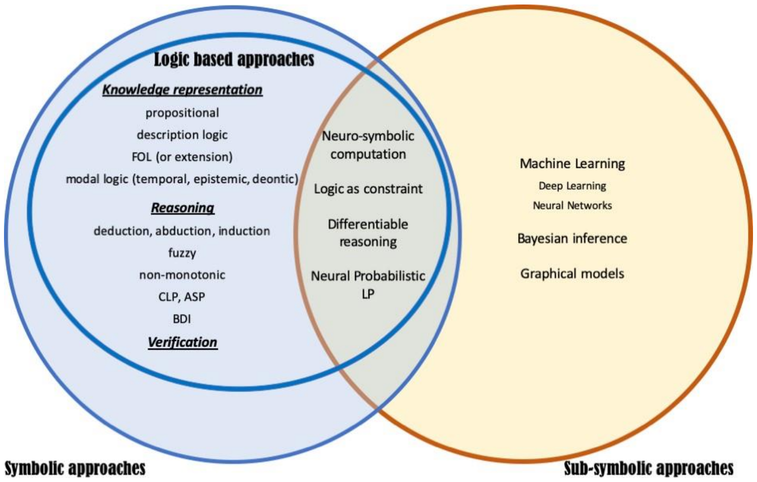


Figure 2.1.1. Classification of logic-based technologies

To give a better idea of how much the *symbolic* approach is used today, Table 2.1.1. illustrates the sorts of logic per application area[29]:

	FOL	DL	EL	TL	FL	PL	DR	CLP
Formalization & Verification				C		C		
Cognitive Agents	C	C	C	C	C		C	
Healthcare & Wellbeing	C	C				C		
Law & Governance	C						C	
Education	C	C			C			
Planning & Task Allocation	C	C					C	C
Robotics		C		C	C			C

Table 2.1.1. Applications of logic to technology

[28] Figure taken from Calegari *et al.* (2020).

[29] Acronym and abbreviation key: FOL: First-Order Logic; DL: Description Logic; EL: Epistemic Logic; TL: Temporal Logic; FL: Fuzzy Logic; PL: Probabilistic Logic; DR: Defeasible Reasoning; CLP: Constraint Logic Programming.

Table 2.1.1. shows that cognitive agents and robotics are the application areas that exploit the widest spectrum of logic-based approaches. Reading the table orthogonally, first-order logic and description logic appear to be the most widely used.

Overall, the symbolic approaches appears to be a viable and promising option to face key issues in today's intelligent systems.

2.2. Propositional Logic: Syntax and Semantics

Modern logic is a formal, symbolic system that tries *to discern the laws of truth*[30]. As Gottlob Frege, one of the founders of modern logic, put it:

> Just as "beautiful" points the ways for aesthetics and "good" for ethics, so do words like "true" for logic. All sciences have truth as their goal; but logic is also concerned with it in a quite different way: logic has much the same relation to truth as physics has to weight or heat. To discover truths is the task of all sciences; it falls to logic to discern the laws of truth[31].

To define this logic, a (countably infinite) set of propositions $\text{Prop} = \{p_i | i \in \mathbb{N}\}$ will be assumed. The formulas of propositional logic (PL) will be strings over the alphabet $\text{Prop} \cup \{(,), \neg, \wedge, \vee, \rightarrow, \bot\}$[32].

[30] Smith (2012). The origins of the classical propositional logic, as it was, and still often is called, go back to antiquity and are due to giants of Western thought such as Plato (1921) and Aristotle (1923, 1963, 1975) and to the Stoic school of philosophy, whose most eminent representative was Chryssipus. But the real development of this calculus began only in the mid-19th century and was initiated by the research done by the English mathematician Boole (1847), who is sometimes regarded as the founder of mathematical logic. The classical propositional calculus was first formulated as a formal axiomatic system by the eminent German logician Frege (1879).

[31] Frege (1918-19, p. 351).

[32] In general, the job of describing a logical system comes in three parts: grammar, semantics and proofs. Grammar describing what counts as a

Definition 2.2.1. The set of *well formed formulas* (wff) in propositional logic is the smallest set satisfying the following properties:

- $\bot$ is a wff.
- Any proposition p_i (by itself) is a wff.
- If φ is wff then $\neg\varphi$ is a wff.
- If φ and ψ are wffs then $\varphi \wedge \psi$ is a wff.
- If φ and ψ are wffs then $\varphi \vee \psi$ is a wff.
- If φ and ψ are wffs then $\varphi \rightarrow \psi$ is a wff.

wffs φ in propositional logic are given by the following *BNF grammar*:

$$\varphi ::= p|\bot|\neg\varphi|\varphi \wedge \varphi|\varphi \vee \varphi|\varphi \rightarrow \varphi$$

Where p is an element of Prop.

Our semantic will follow the inductive definition of the syntax. The semantics of formulas in a logic, are typically defined with respect to a model, which identifies a “world” in which certain facts are true. In the case of propositional logic, this world or model is a truth valuation or assignment that assigns a truth value (true/false) to every proposition. The *true value* will be denoted by 1, and the *false value* will be denoted by 0.

Definition 2.2.2. A truth valuation or assignment is a function $\mathcal{V}$ that assigns truth values to each of the propositions, i.e., $\mathcal{V}$: Prop $\rightarrow \{1,0\}$. The value of a proposition p under valuation $\mathcal{V}$ is given by $\mathcal{V}(p)$. Semantics will be defined through a satisfaction

formula, semantics defining truth in a model and proofs describing what counts as a proof.

relation, which is a binary relation $\vDash$ between valuations and formulas. The statement $\mathcal{V} \vDash \varphi$ should be read as “$\mathcal{V}$ satisfies "’ or “φ is true in $\mathcal{V}$” or “$\mathcal{V}$ is a model of φ”. It is defined inductively following the syntax of formulas. In the definition below, it is said $\mathcal{V} \nvDash \varphi$ when $\mathcal{V} \vDash \varphi$ does not hold.

Definition 2.2.3. For a valuation $\mathcal{V}$ and wff φ, the satisfaction relation, $\mathcal{V} \vDash \varphi$, is defined inductively based on the structure of φ as follows:

- $\mathcal{V} \vDash p$ iff $\mathcal{V}(p) = 1$.

- $\mathcal{V} \vDash \bot$ is never true. That is, $\mathcal{V} \nvDash \bot$.

- $\mathcal{V} \vDash \neg\varphi$ iff $\mathcal{V} \nvDash \varphi$.

- $\mathcal{V} \vDash \varphi \wedge \psi$ iff $\mathcal{V} \vDash \varphi$ and $\mathcal{V} \vDash \psi$.

- $\mathcal{V} \vDash \varphi \vee \psi$ iff $\mathcal{V} \vDash \varphi$ or $\mathcal{V} \vDash \psi$.

- $\mathcal{V} \vDash \varphi \to \psi$ iff either $\mathcal{V} \vDash \psi$ or $\mathcal{V} \nvDash \varphi$.

Example 2.2.1. A couple of examples to understand how the inductive definition of the satisfaction relation can be applied. Consider the formula $\varphi = \big(r \to (q \to p) \wedge \big(q \vee (r \to p)\big)\big)$. Consider the valuation $\mathcal{V}_1$ that sets all propositions to 1. Now $\mathcal{V}_1 \vDash \varphi$ can be seen from the following observations:

$\mathcal{V}_1 \vDash p$	because $\mathcal{V}_1(p) = 1$
$\mathcal{V}_1 \vDash q \to p$	semantics of $\to$
$\mathcal{V}_1 \vDash r \to (q \to p)$	semantics of $\to$
$\mathcal{V}_1 \vDash r \to p$	semantics of $\to$
$\mathcal{V}_1 \vDash q \vee (r \to p)$	semantics of $\vee$

$\mathcal{V}_1 \vDash \left(r \to (q \to p) \wedge \left(q \vee (r \to p)\right)\right)$ semantics of $\wedge$

Consider $\mathcal{V}_2$ that assigns all propositions to 0. Once again $\mathcal{V}_2 \vDash \varphi$. The reasoning behind this observation is as follows.

$\mathcal{V}_1 \nvDash r$ because $\mathcal{V}_1(r) = 0$

$\mathcal{V}_1 \vDash r \to (q \to p)$ semantics of $\to$

$\mathcal{V}_1 \vDash r \to p$ semantics of $\to$

$\mathcal{V}_1 \vDash q \vee (r \to p)$ semantics of $\vee$

$\mathcal{V}_1 \vDash \left(r \to (q \to p) \wedge \left(q \vee (r \to p)\right)\right)$ semantics of $\wedge$

The semantics in Definition 2.2.3. defines a satisfaction relation between valuations and formulas. However, one could define the semantics of propositional logic differently, by considering the formula as a “program” or “circuit” that computes a truth value based on the assignment. This approach is captured by the following definition of the value of a wff under a valuation.

Definition 2.2.4. The value of a wff φ under valuation $\mathcal{V}$, denoted by $\mathcal{V}[\![\varphi]\!]$, is inductively defined as follows:

- $\mathcal{V}[\![\bot]\!] = 0$

- $\mathcal{V}[\![p]\!] = \mathcal{V}(p)$

- $\mathcal{V}[\![\neg\varphi]\!] = \begin{cases} 1 \text{ if } \mathcal{V}[\![\varphi]\!] = 0 \\ 0 \text{ if } \mathcal{V}[\![\varphi]\!] = 1 \end{cases}$

- $\mathcal{V}[\![\varphi \wedge \psi]\!] = \begin{cases} 1 \text{ if } \mathcal{V}[\![\varphi]\!] = 1 \text{ and } \mathcal{V}[\![\psi]\!] = 1 \\ 0 \text{ otherwise} \end{cases}$

- $\mathcal{V}[\![\varphi \vee \psi]\!] = \begin{cases} 0 \text{ if } \mathcal{V}[\![\varphi]\!] = 0 \text{ and } \mathcal{V}[\![\psi]\!] = 0 \\ 1 \text{ otherwise} \end{cases}$

- $\mathcal{V}[\![\varphi \to \psi]\!] = \begin{cases} 0 \text{ if } \mathcal{V}[\![\varphi]\!] = 1 \text{ and } \mathcal{V}[\![\psi]\!] = 0 \\ 1 \text{ otherwise} \end{cases}$

Example 2.2.2. Let us consider $\varphi = \Big(r \to (q \to p) \wedge \big(q \vee (r \to p)\big)\Big)$ and $\mathcal{V}_1$ which assigns all propositions to 1, from Example 1. $\mathcal{V}[\![\varphi]\!]$ can be computed as follows:

$\mathcal{V}_1[\![p]\!] = 1$	because $\mathcal{V}_1(p) = 1$
$\mathcal{V}_1[\![q \to p]\!] = 1$	semantics of →
$\mathcal{V}_1[\![r \to (q \to p)]\!] = 1$	semantics of →
$\mathcal{V}_1[\![r \to p]\!] = 1$	semantics of →
$\mathcal{V}_1[\![q \vee (r \to p)]\!] = 1$	semantics of ∨
$\mathcal{V}_1\left[\!\!\left[\Big(r \to (q \to p) \wedge \big(q \vee (r \to p)\big)\Big)\right]\!\!\right] = 1$	semantics of ∧

Definitions 2.2.2. and 2.2.4. are both equivalent in some sense. This is captured by the following theorem:

Theorem 2.2.1. For any truth valuation $\mathcal{V}$ and wff φ, $\mathcal{V} \vDash \varphi$ *if and only if* $\mathcal{V}[\![\varphi]\!] = 1$.

Definition 2.2.5. The model of wff φ is the set of valuations that satisfy φ. More precisely:

$$[\![\varphi]\!] = \{\mathcal{V} | \mathcal{V} \vDash \varphi\}$$

Definition 2.2.6. (Logical Equivalence). A wff φ is said to be *logically equivalent* to ψ iff any of the following equivalent conditions hold.

- for every valuation $\mathcal{V}$, $\mathcal{V} \vDash \varphi$ iff $\mathcal{V} \vDash \psi$,
- for every valuation $\mathcal{V}$, $\mathcal{V}[\![\varphi]\!] = \mathcal{V}[\![\psi]\!]$
- $\mathcal{V}[\![\varphi]\!] = \mathcal{V}[\![\psi]\!]$.

Definition 2.2.7. (Logical Consequence). Let Γ be a (possibly infinite) set of formulas and let φ be a wff. It is said that $\Gamma \vDash \varphi$ iff $\bigcap_{\psi \in \Gamma} [\![\varphi]\!] \subseteq [\![\psi]\!]$.

Definition 2.2.8. (Tautologies). A wff φ is a *tautology* or is *valid* if for every valuation $\mathcal{V}, \mathcal{V} \vDash \varphi$.

Definition 2.2.9. (Satisfiable). A formula φ is *satisfiable* if there is some valuation $\mathcal{V}$ such that $\mathcal{V} \vDash \varphi$. In other words, $[\![\varphi]\!] \neq \emptyset$.

2.3. Semantic Tableaux as a Decision Procedure

Historically, a mathematical problem is considered “closed” when a proof, or better still an algorithm, is found to solve it “in principle”. In this sense the deducibility problem of classical propositional logic was already “closed” in the early 1920’s, when

Wittgenstein[33] and Post[34] independently devised the well-known *decision procedure* based on the truth-tables.

$\boldsymbol{\varphi}$	$\boldsymbol{\neg\varphi}$
1	0
0	1

$\boldsymbol{\varphi}$	$\boldsymbol{\psi}$	$\boldsymbol{\varphi \wedge \psi}$
1	1	1
1	0	0
0	1	0
0	0	0

$\boldsymbol{\varphi}$	$\boldsymbol{\psi}$	$\boldsymbol{\varphi \rightarrow \psi}$
1	1	1
1	0	0
0	1	1
0	0	1

$\boldsymbol{\varphi}$	$\boldsymbol{\psi}$	$\boldsymbol{\varphi \vee \psi}$
1	1	1
1	0	1
0	1	1
0	0	0

Table 2.3.1. Truth-tables

A proof is a mechanism for showing that a given claim ψ is a logical consequence of some premises $\varphi_1, \dots, \varphi_k$. In this view, the purpose of a proof is to make explicit what is already implicitly present. A proof is presented as a finite sequence of steps, each of which is either an axiom or the logical conclusion of a set of steps occurring earlier in the proof. The final step of the proof is the demonstration of the truth of the claim ψ. A formal proof requires that all implicit assumptions are made explicit and the steps in the proof are shown with reference to the sources used in deriving each step. The method of semantic tableaux and their formal presentation in propositional logic framework are discussed below.

[33] Wittgenstein (1921).

[34] Post (1921).

The theoretical foundations of proof by contradiction (or *reductio ad absurdum*) will now be investigated. Frequently, rather than make a “direct” proof of a formula φ, it is easier to start by assuming $\varphi = 0$ and proving from that a contradiction. This method of proof is an application of the rule of logical reasoning known as *modus tollens*. According to this rule, a proposition is proved by showing that its falseness leads to unacceptable consequences. Wishing to prove a formula φ, one first assumes its negation $\varphi = 0$ to be true. One then goes on to show that φ implies ψ, where ψ is already known to be false. By this argument, $\varphi = 0$ must be false and it follows logically that its negation, the original proposition φ must be true. To reiterate, the aim of a proof by contradiction is essentially to contradict one of our assumptions. If the aim is achieved then, as a consequence of modus tollens, the original statement must be true. Figure 2.3.1. proposes a general model for proof by contradiction:

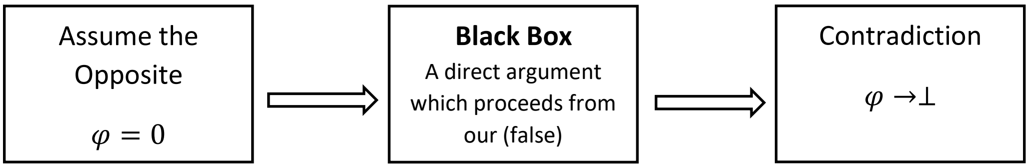

Assume the Opposite
$\varphi = 0$
Black Box
A direct argument which proceeds from our (false)
Contradiction
$\varphi \rightarrow \perp$

Figure 2.3.1. Model for proof by contradiction

On these theoretical grounds, the method of semantic or analytic tableaux is a refutation procedure for proving theorems of a logic. The procedure has the intention of proving the satisfiability of given formula by constructing a tableau. For our purposes it suffices to visualize tableaux as binary trees. The branches of a tableau are implicitly disjunctively connected, and the formulae on a branch are implicitly conjunctively connected. A branch is identified with the set of formulae it contains. To every formula that is not a literal exactly one tableau expansion rule can be applied. Following Smullyan[35], formulae for PL are divided into two classes with corresponding rules, namely α (conjunctive propositional) and β (disjunctive propositional). The formula classes are summarized in Table 2.3.2. and Table 2.3.3. shows the expansion rule schema for the ground version of tableaux.

α	α_1	α_2
$\varphi \wedge \psi = 1$	$\varphi = 1$	$\psi = 1$
$\varphi \vee \psi = 0$	$\varphi = 0$	$\psi = 0$
$\varphi \rightarrow \psi = 0$	$\varphi = 1$	$\psi = 1$
$\neg\varphi = 1$	$\varphi = 0$	
$\neg\varphi = 0$	$\varphi = 1$	

β	β_1	β_2
$\varphi \vee \psi = 1$	$\varphi = 1$	$\psi = 1$
$\varphi \wedge \psi = 0$	$\varphi = 0$	$\psi = 0$
$\varphi \rightarrow \psi = 1$	$\varphi = 0$	$\psi = 1$

Table 2.3.2. Formula classes

$$\frac{\alpha}{\begin{array}{c}\alpha_1\\ \alpha_2\end{array}} \qquad \frac{\beta}{\beta_1 \mid \beta_2}$$

Table 2.3.3. Tableaux expansion rule schema

A tableau T is expanded by choosing a branch B of T and a formula $\varphi \in B$ and extending B by as many subbranches as the rule corresponding to φ has extensions; the new subbranches contain the formulae in the extensions.

[35] Smullyan (1978).

In this way, a set of branches is obtained, each populated by ever smaller formulas whose truth or falsehood follows from the truth or falsehood of some formulas occurring earlier on the branch. When a direct contradiction is found between these formulas, the relevant branch is closed. As long as the branch remains open, it can be thought of as corresponding to the set of all models that satisfy its formulas. If the rules for expanding a branch are exhausted without running into any contradiction, then the idea is that there must be a model that satisfies the formulas. The proof of the validity of φ then amounts to a refutation of $\varphi \rightarrow \bot$. In other words, to prove a theorem, it is necessary to close every branch that contains its negation, to show that there is no counter-model to the theorem.

Therefore, to prove a sentence φ to be a tautology, expansion rules are applied starting from the initial tableaux that consists of the single node $\varphi = 0$. A proof is found, if all branches of the constructed tableau are closed (contain complementary formulae):

Theorem 2.3.1. A propositional sentence φ is a tautology if and only if there is a sequence $T_0, \dots, T_n$ of tableaux $(n \geq 0)$ such that

1. T_0 consists of the single node $\varphi = 0$.
2. For $1 \leq i \leq n$ the tableau T_i is constructed from T_{i-1} by applying one of the tableau expansion rules from Table 2.3.3.
3. All branches of T_n are closed, i.e., contain complementary formulae $\varphi = 1$ and $\varphi = 0$.

The construction of a closed tableau is a highly indeterministic process, because at each step one is free to choose a branch B of the tableau and a formula $\varphi \in B$ for expansion.

Example 2.3.1. Let us consider the first law of contraposition $\varphi = (\neg p \to \neg q) \to (q \to p)$

$$
\begin{array}{cc}
(\neg p \to \neg q) \to (q \to p) = 0 & \\
| & \\
\neg p \to \neg q = 1 & \alpha \to \\
| & \\
q \to p = 0 & \alpha \to \\
| & \\
q = 1 & \alpha \to \\
| & \\
p = 0 & \alpha \to \\
/ \quad \backslash & \\
q = 0 \quad\quad p = 1 & \beta \to \\
\otimes \quad\quad\quad \otimes &
\end{array}
$$

Since all branches of the constructed tableaux are closed, φ is a tautology.

On the basis of the proof by contradiction and the tableaux rules, it is possible to simplify the tableau calculus, and thus optimise the concept and the schema of the decision procedure that will be used in the next chapter[36]. The procedure, in contrast to the rules

[36] It is important to specify that this dissertation is inspired by the method and logical rigor of *proof theorists*. On proof theory see Galvan et al. (2021). Furthermore, this book allowed Sergio Galvan, Paolo Mancosu and Richard Zach to win *Shoenfield Prize Recipients* (“Nobel Prize” in the field of logic).

developed by Smullyan, does not distinguish between *alpha* or *beta* rules, but is based on a single class. Each logical connective is associated with a precise rule and, with the exception of negation, which is a unary connective, for the other operators a distinction is made between the first conjunction/disjunction or antecedent and the second conjunction/disjunction or consequent. The antecedent will be indicated in subscript with α, while the consequent will be indicated with β.

Furthermore, will be denoted by superscript with $[1/0]$ the fact that a proposition φ is true $[1]$ or false$[0]$. The *decision rules* are as follows:

- $\langle \neg\varphi^{[1/0]} \rangle: \begin{cases} \langle \neg\varphi^{[1]} \rangle: \varphi = 0 \text{ iff } \varphi = 1 \\ \langle \neg\varphi^{[0]} \rangle: \varphi = 1 \text{ iff } \varphi = 0 \end{cases}$

- $\langle \wedge\, \varphi_{\alpha/\beta}^{[1/0]} \rangle: \begin{cases} \langle \wedge\, \varphi_{\alpha}^{[1]} \rangle: \varphi_{\alpha} = 1 \text{ iff } \varphi = 1 \\ \langle \wedge\, \varphi_{\alpha}^{[0]} \rangle: \varphi_{\alpha} = 0 \text{ iff } \varphi = 0 \\ \langle \wedge\, \varphi_{\beta}^{[1]} \rangle: \varphi_{\beta} = 1 \text{ iff } \varphi = 1 \\ \langle \wedge\, \varphi_{\beta}^{[0]} \rangle: \varphi_{\beta} = 0 \text{ iff } \varphi = 0 \end{cases}$

- $\langle \vee\, \varphi_{\alpha/\beta}^{[1/0]} \rangle: \begin{cases} \langle \vee\, \varphi_{\alpha}^{[1]} \rangle: \varphi_{\alpha} = 1 \text{ iff } \varphi = 1 \\ \langle \vee\, \varphi_{\alpha}^{[0]} \rangle: \varphi_{\alpha} = 0 \text{ iff } \varphi = 0 \\ \langle \vee\, \varphi_{\beta}^{[1]} \rangle: \varphi_{\beta} = 1 \text{ iff } \varphi = 1 \\ \langle \vee\, \varphi_{\beta}^{[0]} \rangle: \varphi_{\beta} = 0 \text{ iff } \varphi = 0 \end{cases}$

- $$\langle \to \varphi_{\alpha/\beta}^{[1/0]} \rangle : \begin{cases} \langle \to \varphi_{\alpha}^{[1]} \rangle : \varphi_{\alpha} = 0 \text{ iff } \varphi = 1 \\ \langle \to \varphi_{\alpha}^{[0]} \rangle : \varphi_{\alpha} = 1 \text{ iff } \varphi = 0 \\ \langle \to \varphi_{\beta}^{[1]} \rangle : \varphi_{\beta} = 1 \text{ iff } \varphi = 1 \\ \langle \to \varphi_{\beta}^{[0]} \rangle : \varphi_{\beta} = 0 \text{ iff } \varphi = 0 \end{cases}$$

Starting from the proof by absurdity, the rules make it possible to create an algorithm capable of determining whether a formula φ is a tautology. The defined procedure supposes that the propositional form is false and one proceeds until one encounters a possible contradiction, i.e. to deduce both a proposition and its negation; if this happens, one can conclude that the propositional form, since it cannot be false, is a tautology.

A tableaux procedure P is expanded by choosing a decision sequence D of P and a formula $\varphi \in D$ and extending D by as many sub-sequences as there are extensions of the rule corresponding to φ; the new sub-sequences contain the formulas of the extensions.

Theorem 2.3.2. A propositional sentence φ is a tautology if and only if there is a sequential procedure $P_0, \dots, P_n$ $(n \geq 0)$ such that

1. P_o consists of the single node $\varphi = 0$.
2. For $1 \leq i \leq n$ the sequential procedure P_i is constructed from P_{i-1} by applying one of decision rules.
3. All sub-sequences of P_n are closed, i.e., contain complementary formulae $\varphi = 1$ and $\varphi = 0$.

The procedure can be illustrated with two examples.

Example 2.3.2. The first example considers the most famous principle of logic, first formulated by Aristotle in *Metaphysics* IV, the principle of non-contradiction[37] $\varphi = \neg(p \wedge \neg p)$

1. $\neg(p \wedge \neg p) = 0$
2. $p \wedge \neg p = 1$ $\quad$ $(1)\ \langle \neg\varphi^{[0]} \rangle$
3. $p = 1$ $\quad$ $(2)\ \langle \wedge\ \varphi_{\alpha}^{[1]} \rangle$
4. $\neg p = 1$ $\quad$ $(2)\ \langle \wedge\ \varphi_{\beta}^{[1]} \rangle$
5. $p = 0$ $\quad$ $(4)\ \langle \neg\varphi^{[1]} \rangle$

Example 2.3.3. The second example considers the principle of Pseudo-Scotus[38] $\varphi = p \wedge \neg p \to q$

1. $p \wedge \neg p \to q = 0$
2. $p \wedge \neg p = 1$ $\quad$ $(1)\ \langle \to\ \varphi_{\alpha}^{[0]} \rangle$
3. $q = 0$ $\quad$ $(1)\ \langle \to\ \varphi_{\beta}^{[0]} \rangle$
4. $p = 1$ $\quad$ $(2)\ \langle \wedge\ \varphi_{\alpha}^{[1]} \rangle$
5. $\neg p = 1$ $\quad$ $(2)\ \langle \wedge\ \varphi_{\beta}^{[1]} \rangle$
6. $p = 0$ $\quad$ $(5)\ \langle \neg\varphi^{[1]} \rangle$

[37] There are also snippets of discussion about the principle of non-contradiction early in the corpus, for example in *De Interpretatione* (1963), and there is the chapter 11 of *Posterior Analytics* I (1975), but none of these rival Aristotle's treatment of the principle of non-contradiction in *Metaphysics* IV 3–6, especially 4 (1923).

[38] In classical and intuitionistic logic, the principle of Pseudo-Scotus or the principle of explosion *ex falso [sequitur] quodlibet* (from falsehood, anything [follows]), or *ex contradictione [sequitur] quodlibet* (from contradiction, anything [follows]), is the law according to which any statement can be proven from a contradiction.

Of course, as is clearly visible in the examples, also in our optimization the validity of tableaux as a proof calculus depends on if it is both sound, i.e., proves tautologies only, and complete, i.e., proves all tautologies.

Soundness. The proposition that φ is a tautology is equivalent to $\neg\varphi$ being false under any $\mathcal{V}$ (hence: is not satisfiable). A tableau succeeds in showing the latter as it only closes if a contradiction is found on every branch.

Completeness. Completeness for tableau can be proven using contraposition "φ has no corresponding closed tableau $\Longrightarrow$ φ is no tautology" and is largely based on the concept of consistency. Therefore, a formula without a closed tableau is called tableau consistent which in particular means that the formula is satisfiable.

2.4. Predicate Logic

A predicate is a way of indicating that a certain variable has a "property" that characterizes it. For example, suppose that x has the characteristic "to be a car", that is to say x is a car. Well, a way to formalize it would be, for example: $C(x)$. This $C(x)$ is what is called a predicate. However, in natural language, one often wants to say how many objects have a particular property. For example, one can say: "There are six children in the room", or "Most people are polite" or "Everyone has a mother", or "Something is coming up", or "Nothing is impossible". Words like "most", "all", "none" and "some" are called quantifiers because they tell us how many objects have a certain property. Predicate logic, or first-order logic, has two quantifiers: "all" and "at least one." These two quantifiers are represented using the symbols ∀ and ∃ respectively, as shown in the following table 2.4.1.

Symbol	Name of Quantifier	English Equivalent
∀	Universal quantifier	All
∃	Existential quantifier	At least one

Table 2.4.1. Table of quantifiers

Syntax and semantics of first-order logic are characterised as follows. Syntactic rules to specify the wffs of predicate logic:

a. If t_1, t_2, t_n are individual terms and P is an n-place predicate, then $P(t_1, t_2, t_n)$ is a wff.

b. If x is an individual variable and φ is a wff, then $\exists x\varphi$ and $\forall x\varphi$ are wffs.

c. If φ and ψ are wffs, then $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$, and $(\varphi \leftrightarrow \psi)$ are wffs

d. Only the formulas generated in accordance with these rules are wffs

As is well known, in propositional logic, the semantic meaning of an expression is a function that takes a “truth assignment” (assignment of truth values 0 and 1 to the propositional variables of E) as an argument and produces 0 or 1 as a result. The result is determined by the evaluation of E with the atomic operands replaced by 0 or 1, according to the given truth assignment. A truth assignment, in turn, is a function that takes propositional variables p as arguments and returns 0 or 1 for each.

$$p \qquad \xrightarrow[\substack{Truh \\ assignment}]{} \qquad 0 \text{ or } 1$$

Alternatively, it is possible to see a truth assignment as a table that gives, for each propositional variable, a truth value, 0 or 1.

$Truth$ $\overrightarrow{Meaning}$ 0 or 1
$assignment$

In predicate logic, it is not sufficient to assign a constant 0 or 1. More precisely, one must first pick a nonempty domain D of values, from which the values of the variables can be selected[39]. Will be hired, for convenience, that the domain includes any constants appearing in the expression itself.

Now, let P be a predicate with k arguments. Then an interpretation for predicate P is a function that takes as input an assignment of domain elements to each of the k arguments of P and returns 0 or 1[40]. In this way, the meaning of expressions in predicate logic will be defined respectively:

1. A nonempty domain D, including any constants appearing in E
2. An interpretation for each predicate P appearing in E, and
3. A value in D for each of the free variables of E, if any

$values\ for$ $\Rightarrow$ $interpreation$ $\Rightarrow$ 0 or 1
$argument$ $for\ a\ predicate$

[39] In general, this domain could be anything: reals, integers, or some set of values with no particular name or significance.

[40] Equivalently, the interpretation of P can be seen as a relation with k columns. For each assignment of values to the arguments that makes P true in this interpretation, there is a tuple of the relation.

2.5. Modal Logic

Modal logics are *extensions* of classical logic[41]. Like classical logic, modal logic was first discussed in a systematic way by Aristotle in *De Interpretatione*. Philosophers after Aristotle added other interesting observations regarding modal reasoning. Contributions were made by the Megarians, the Stoics, Ockham, and Pseudo-Scotus, among others. Before the last century, work in modal logic after the Scholastics was practically non-existent, except for the intuition of Leibniz who suggests that there are other possible worlds besides the actual world. However, the innovations in modal logic that will be used in this dissertation were developed by S. Kripke[42], although they were anticipated by the work of S. Kanger[43] and J. Hintikka[44].

Strictly speaking, modal logic is the study of modal propositions and the logical relationships that they bear to one another. Modal propositions contain expressions such as “it is necessary that” and “it is possible that”. For example, the following are all modal propositions:

It is possible that it will snow tomorrow.

It is not possible for humans to live on Neptune.

It is necessary that either it is raining here now or it is not raining here now.

[41] On modal logic and, in general, intensional logics see Galvan (1991); Frixione *et al*. (2016); Turbanti (2020).

[42] Kripke (1959).

[43] Kanger (1957).

[44] Hintikka (1957, 1961, 1962).

A proposition p is not possible if and only if the negation of p is necessary.

The operators "it is possible that" ◇ (diamond) and "it is necessary that" □ (box) are called "modal" operators because they specify a mode in which the rest of the proposition can be said to be true.

To keep things simple, a propositional language $\mathcal{L}$ will be used, containing certain "proposition letters" as atomic sentences $AT: p, q, r, \dots$. The operators of classical logic will be present $\neg, \wedge, \vee, \rightarrow$, and the box □ and diamond ◇ of necessity and possibility. Will be used $\varphi, \psi, \dots$ as metavariables for formulas of $\mathcal{L}$. Further expressions will be constructed inductively using the following format:

$$\varphi \ ::= AT|\neg\varphi|\varphi \wedge \psi|\varphi \vee \psi|\varphi \rightarrow \psi|\Diamond\varphi| \Box \varphi$$

The next step is to define models or interpretations for the language $\mathcal{L}$. A *normal possible worlds frame* or *Kripke frame* $\mathcal{F}$ for $\mathcal{L}$ is a pair $\langle W, R\rangle$ where W is a set of possible worlds and $R \subseteq W \times W$ is a binary accessibility relation between them. A Kripke frame with an evaluation function v is called a *model* $\mathcal{M} = \langle W, R, v\rangle$[45]. This assigns to each atom either the value 1 or the value 0 at a world. In this way "$v_w(p) = 1$" means that p is true at w, and "$v_w(p) = 0$" means that it is false there.

The evaluation function v is extended to the entire language through the following recursive clauses:

(S¬) $v_w(\neg\varphi) = 1$ if $v_w(\varphi) = 0$, and 0 otherwise.

[45] These models allow one to define the model-theoretic notions of truth, logical truth, and logical consequence.

$(\mathrm{S}\wedge)\ v_w(\varphi \wedge \psi) = 1$ if $v_w(\varphi) = 1$ and if $v_w(\psi) = 1$, and 0 otherwise.

$(\mathrm{S}\vee)\ v_w(\varphi \vee \psi) = 1$ if $v_w(\varphi) = 1$ or if $v_w(\psi) = 1$, and 0 otherwise.

$(\mathrm{S} \rightarrow)\ v_w(\varphi \rightarrow \psi) = 1$ if $v_w(\varphi) = 0$ or if $v_w(\psi) = 1$, and 0 otherwise.

$(\mathrm{S}\Diamond)\ v_w(\Diamond\varphi) = 1$ if for some $w_1 \in W\ such\ that\ Rww_1,$

$v_{w_1}(\varphi) = 1,$ and 0 otherwise.

$(\mathrm{S}\ \Box)\ v_w(\Box\ \varphi) = 1$ if for all $w_1 \in W\ such\ that\ Rww_1,$

$v_{w_1}(\varphi) = 1,$ and 0 otherwise.

Where Rww_1 indicates the *accessibility relation*. Kripke's idea here is that not every world is modally accessible from a given world w. A world w can access a world w_1 (or, conversely, w_1 is accessible from w) just in case every proposition that is true at w_1 is possibly true at w. Kripke's definition was:

> "Necessarily p" is true at a world w if and only if p is true at every world w_1 accessible from w.

Therefore, a sentence "necessarily p" is true at world w so long as p is true at all the worlds that are possible from the point of view of w.

Logical consequence "⊨", is defined as truth preservation at all worlds of all models (for any set of formulas Δ):

> $\Delta \vDash \varphi$ iff for all models $\mathcal{M} = \langle W, R, v\rangle$ and all $w \in W$: if $v_w(\varphi) = 1$ for all $\psi \in \Delta$, then $v_w(\varphi) = 1$.

All modal calculations have one rule in common, the Necessitation rule:

(N) if $\vdash \varphi,$ then $\vdash\Box\ \varphi$

The logic induced by the semantics is called **K**, after Kripke[46].

(K) $\Box (\varphi \rightarrow \psi) \rightarrow (\Box \varphi \rightarrow \Box \psi)$

This is the weakest *normal modal logic*. Within modal logic, "normal" means that the logic includes all the classical tautologies plus (K), and is closed under *modus ponens* and the *necessitation rule.*

Moreover, **K** is the basic normal modal logic, since its semantics does not place conditions on the accessibility relation R. Given some precise conditions on R, stronger normal modal logics will be obtained. The stronger normal modal logics obtained in this way contain all the **K**-theorems, plus some extra ones too. Table 2.5.1. shows the well-known axioms:

Axiom name	Axiom scheme	Frame condition
D	$\Box \varphi \rightarrow \Diamond\varphi$	R is serial: $\forall x \exists y Rxy$
T	$\Box \varphi \rightarrow \varphi$	R is reflexive: $\forall x Rxx$
B	$\varphi \rightarrow \Box \Diamond \varphi$	R is symmetrical: $\forall x \forall y (Rxy \rightarrow Ryx)$
4	$\Box \varphi \rightarrow \Box \Box \varphi$	R is transitive: $\forall x \forall y \forall z (Rxy \land Ryz \rightarrow Rxz)$
5	$\Diamond\varphi \rightarrow \Box \Diamond \varphi$	R is eucledian: $\forall x \forall y \forall z (Rxy \land Rxz \rightarrow Ryz)$

Table 2.5.1. Axioms of modal logic

[46] The semantics makes $\Box \varphi$ equivalent to $\neg\Diamond\neg\varphi$ and $\Diamond\varphi$ equivalent to $\neg \Box \neg\varphi$.

2.6. Epistemic Logic

In contemporary epistemology, it is widely accepted that truth is a necessary condition of knowledge. For this reason: *i*) knowledge is said to be factual, i.e., the truth of the known proposition is presupposed; *ii*) for knowledge to exist, belief must be entertained; *iii*) belief must be justified. For a long time, truth, belief and justification were considered jointly sufficient conditions for knowledge to exist. From the 1960s onwards, thanks to the work of Gettier[47], contemporary epistemologists have argued that, in addition to the three conditions, others are required. However, as much as logicians are particularly interested in the complex debate that has developed among epistemologists concerning the strategy to be adopted to characterize knowledge exhaustively, in epistemic logics knowledge is generally characterised as simple true belief. In this way, the logicians treat knowledge and belief attributions as formulae containing modal operators. Semantically, this means that, when assessing the truth value of a formula associated with an epistemic operator, one considers a set of alternative circumstances.

As seen in the previous section, such alternative circumstances are called possible worlds. Let us suppose for example that a subject believes that Giorgia Meloni is the Italian Prime Minister and that Barack Obama is the President of the United States. The worlds compatible with his beliefs will be all and only worlds in which it is true that Giorgia Meloni is the Prime Minister of Italy and Barack Obama is the President of the United States. But according to the semantics of possible worlds, the

[47] Gettier (1963).

possession or non-possession of knowledge depends on how things are in the actual world: our subject cannot know that Barack Obama is the President of the United States, since this is false.

Based on this, Hintikka[48] provided a semantic interpretation of epistemic and belief operators which can be presented 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 2.6.1. [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 $K_a\varphi$ (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 as will be seen in chapter 4, in our case, a "Metadata extraction agent" (MEA).

To build an interpretation, is first introduced 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 2.6.2. [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

[48] Hintikka (1962).

- $W \neq \emptyset$ 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 $a \in \mathcal{A}$;
- $V^{\mathcal{P}}: W \rightarrow (\mathcal{P} \rightarrow \{true, false\})$ 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 2.6.3. [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 2.6.4. [Axioms and Inference Rules] The proof system of epistemic logic that will be used is axiomatized by using the axiom of **T** and the rule of modus ponens and necessitation. The full system is presented in Table 2.6.1:

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

Table 2.6.1. Axiom of **T**, modus ponens and necessitation

The reflexivity of R guarantees that the principle

T $K_a\varphi \rightarrow \varphi$

is valid.

Chapter 3

Decision-Making Ontology for a Framework for Metadata Extraction

3.1. Building a Decision-Making Ontology

Decision has inspired reflection of many thinkers since the ancient times. The great philosophers Plato, Aristotle, Thomas Aquinas, René Descartes, Immanuel Kant, Gottlob Frege and Ludwig Wittgenstein, to mention only a few names, reflected, debated and proposed solutions to specific problems. This capacity to pose and solve problems corresponds to the human ability to make *rational decisions*.

In general, a decision is an intellectual act initiated to realise a purpose and a judgement on potential decisions to prescribe a final action. Bernard Roy defines three basic concepts that play a fundamental role in analysing decisions[49]: decision problem, alternatives (potential actions), and criteria.

The decision problem can be characterized by the result expected from a decision-making. In our case, since it is a *choice problem*, the result consists in a subset of potential alternatives. The concept of *alternative* designates the decision object. Finally, a *criterion* can be any kind of information that allows alternatives to be evaluated and compared.

[49] Roy (2005).

Herbert Simon (Nobel Prize in Economics in 1978) was the first to formalize the decision-making process. He proposed a model comprising three main phases: intelligence, design and choice (I.D.C. model)[50]. Intelligence deals with examining an environment for conditions that call for decisions. Planning means developing and analyzing possible decision alternatives. Choice requires selecting an alternative from among those possible.

This process was adapted and extended in different ways. Presently, the commonly agreed and applied decision-making steps are defined as follows:

- define problem,
- identify problem parameters (for instance, alternatives and criteria),
- establish evaluation matrix,
- select method for decision making,
- aggregate evaluations.

These first notions of decision problems allow us to define decision-making (DM) as the result of a cognitive process that leads to the selection of an action among several alternatives. It can be considered as a problem-solving activity that ends when a satisfactory solution is found. As far as knowledge engineering (KE) methodologies are concerned, the topic of DM has already been explored in relation to requirements engineering[51], to methods engineering[52] and, more generally, to systems engineering[53]. Ruhe emphasizes the importance of DM in the field

[50] Simon (1960).
[51] Ngo-The, Ruhe (2005).
[52] Aydin (2006); Kornyshova *et al*. (2007).
[53] Ruhe (2003).

of KE because of: (i) time, effort, quality, and resource constraints; (ii) presence of multiple objectives; (iii) uncertain, incomplete and fuzzy information; and (iv) complex decision space.

In this chapter, a decision-making ontology for a framework for metadata extraction from different document sources will be presented. Our logical-ontological approach, by the name MADME (MAke Decision for Metadata Extraction), will move at the method and model levels. The main objective of the MADME approach is to develop a formal decision-making ontology that can guide the choice of metadata extraction systems.

The MADME approach includes three elements: DM ontology (DMO), DM rules (DMR) and DM procedure (DMP). DMO is an informal and formal representation of digital objects. DMR are derived from DMO and are formal rules written in the language of first-order logic that define all decision steps in detail. The DMP provides a set of methodological guidelines for the application of DMR.

3.2. Informal Decision-Making Ontology

In philosophical contexts, “ontology” has traditionally been defined as the theory of what exist (or of “being *qua* being”): the study of the kind of entities in reality and of the relationships that the entities bear to one another.

As we will regularly use the term “entity” in a broad and generic sense, we here provisionally define it as follows:

> Entity = def. anything that exists, including objects, qualities and processes.

Before analyzing our ontology, it is necessary to present some preliminary metaontological notions from an Aristotelian perspective. According to Schaffer "for Aristotle, metaphysics is about what grounds what[54]". The Aristotelian metaontological paradigm can be characterized as follows:

> Putting this together, the neo-Aristotelian will conceive of the task of metaphysics as: Aristotelian task: The task of metaphysics is to say what grounds what.
>
> That is, the neo-Aristotelian will begin from a hierarchical view of reality ordered by priority in nature. […] The task of metaphysics is to limn this structure. What of the method? A very general answer may be given as:
>
> Aristotelian method: The method of metaphysics is to deploy diagnostics for what is fundamental, together with diagnostics for grounding[55].

In recent times, the use of the term "ontology" has become prominent in computer science and information science and ontologies are designed to promote greater consistency in description of data. Gruber[56] was the first to formulate the term ontology in the field of computer science and defined it as "an explicit specification of a conceptualization". Over the years, numerous approaches have been developed for the creation and application of ontologies based on Gruber's method. For example, Sánchez[57] considers an ontology as a way of representing a

[54] Schaffer (2009, p. 350).

[55] ivi (p. 351).

[56] Gruber (1993).

[57] Sánchez *et al.* (2007).

common understanding of a domain. For Akkermans[58] ontology is a new method for the formation and validation of scientific theories.

The analysis of metaontological and ontological issues is naturally also present in the domain of information system ontologies (ISOs). In general, The Encyclopedia of Database Systems describes ISOs[59] as follows:

> [1] ISOs define a set of representational primitives with which to model a domain (of knowledge). The primitives are typically classes, properties, and relations (among class members). The definitions of such primitives include information about their meaning and constraints on their logically consistent application[60].

Following Tambassi's[61] line of argumentation, [1] has the merit of drawing attention to two focal points:

> [2] the domain to model/systematize;

> [3] the representational primitives for hierarchically and relationally modeling the domain;

In particular, related with [2], another point concerns the aims of ISOs

> [4] ISO denotes an artifact that is *designed* for a purpose;

Such a purpose

[58] Akkermans *et al.* (2006).

[59] For an introduction to the IT debate on ISOs, see Breitman *et al*. (2007); Guarino, Musen (2015).

[60] Cf. Gruber (2009).

[61] Tambassi (2022).

> [5] defines the domain that an ISO aims to represent;
>
> [6] can vary from ISO to ISO;

[5] and [6] allow us to highlight one of the main differences with respect to the aims of philosophical ontology (PO)[62], regarding which the plurality of hypotheses and methods of investigation does not prevent philosophers from arguing that:

> [7] PO aims to study the whole of reality, by providing an (exhaustive) inventory of all there is (or might be)[63];

However, this does not exclude the possibility of regional ontologies, the aim of which is to establish what is within the domain of a specific discipline.

Therefore, while PO's domain concerns in general the whole of reality or some of its specific sub-parts, on the contrary:

> [8] "ISOs' domains are arbitrary: that is, ISOs are in principle open to any domain of knowledge at any level of granularity, as well as being able to deal with anything that each ISO intends to represent[64]".

Our ontology will in some respects come close to both PO and ISO, but in one substantial respect it will differ from both in that our ontology is a decision-making ontology. For this reason, in this dissertation our definition of "ontology" is the following:

Ontology = def. a formal representation, whose representations are intended to designate defined classes, certain relationships between them and specific decision rules.

[62] Runggaldier, Kanzian (1998); Varzi (2011).

[63] Cf. Berto and Plebani (2015).

[64] Tambassi (2022).

According to E.J. Lowe[65], ontological classes or categories are hierarchically organised, although the top-most category must obviously be the most general of all, that of entity or being. At the second-highest level of categorisation all entities are divisible into either universals or particulars. Universals in turn are divisible into properties and relations, and particulars into objects and tropes.

$$\mathit{Entities}\begin{cases}\mathit{Universals}\begin{cases}\mathit{Properties}\\ \mathit{Relations}\end{cases}\\ \mathit{Particulars}\begin{cases}\mathit{Objects}\\ \mathit{Tropes}\end{cases}\end{cases}$$

Given the application of our ontology to heterogeneous document sources, and since our document sources will be digital sources, our ontology, in addition to concrete and abstract objects, will have a third particular category of digital objects.

$$\mathit{Objects}\begin{cases}\mathit{Abstract\ objects}\begin{cases}\mathit{Propositions}\\ \mathit{Sets}\end{cases}\\ \mathit{Concrete\ objects}\begin{cases}\mathit{Masses}\\ \mathit{Organisms}\end{cases}\\ \mathit{Digital\ objects}\begin{cases}\mathit{Documents}\\ \mathit{Extraction\ systems}\\ \mathit{Metadata\ sets}\end{cases}\end{cases}$$

While it is true that there is a vast scientific literature on abstract and concrete objects, it is not easy to provide a precise and rigorous definition of a digital object that would satisfy both the documentation[66] and philosophical sciences.

[65] Lowe (2006).

[66] In particular, in the AgID guidelines on the formation of digital documents, digital objects are defined as: *i*) computerised documents and computerised

For this reason, our reference will be the way the philosopher of information and technology Luciano Floridi characterises digital and analogue predicates

> Both digital and analogue are only "modes of presentation of Being" (to paraphrase Kant), that is, ways in which reality is experienced or conceptualised by an epistemic agent, at a given *level of abstraction* (LoA)[67].

Our *level of abstraction* allows us to identify and define a digital object: *i*) in a metaphysical sense, as "an object composed of a set of bit sequences" (CCSDS, 2012)[68]; and *ii*) in an ontological sense, based on three fundamental classes or categories that cannot be reduced to anything else: (digital) documents, metadata extraction systems and metadata sets.

Documents are divided into four subclasses[69]: text, images, audio, video. The metadata extraction systems class is divided into four subclasses: from text, from images, from audio, from video. For each subclass, there will be specific instances of individuals of extraction systems. The metadata sets class is divided into four subclasses: of text, of images, of audio, of video. Subclasses metadata extraction systems and metadata sets will also have specific instances of individuals[70].

administrative documents with their associated metadata; *ii*) computerised document aggregations with their associated metadata. This interpretation of digital object appears to us to be extremely restrictive both from an ontological and epistemological point of view.

[67] Floridi (2009, p. 152).

[68] Consultative Committee for Space Data Systems (2012).

[69] Subclasses will be indicated using curly brackets.

[70] Instances will be indicated using round brackets.

$$\text{digital objects}\left\{\begin{array}{c}
\mathcal{D}: documents\left\{\begin{array}{c} text: \mathcal{D}_t \\ images: \mathcal{D}_i \\ audio: \mathcal{D}_a \\ video: D_v \end{array}\right. \\
\mathcal{S}: extraction\ systems\left\{\begin{array}{c}
from\ text\left(\begin{array}{c} CERMINE: c \\ OCR++: o \\ GROBID: g \\ FITS: f \\ Apache\ Tika: a \\ EMET: e \end{array}\right. \\
from\ images\left(\begin{array}{c} FITS: f \\ IPTC\ Photo\ Metadata: i \\ Metadata\ extractor: m \\ Apache\ Tika: a \\ EMET: e \end{array}\right. \\
from\ audio\left(\begin{array}{c} FITS: f \\ Apache\ Tika: a \\ EMET: e \end{array}\right. \\
from\ video\left(\begin{array}{c} FITS: f \\ Apache\ Tika: a \\ EMET: e \\ Metadata\ extractor: m \end{array}\right.
\end{array}\right. \\
\mathcal{M}: metadata\ sets\left\{\begin{array}{c}
of\ text\left(\begin{array}{c} CERMINE\ metadata: mc \\ OCR++metadata: mo \\ GROBID\ metadata: mg \\ FITS\ metadata: mf \\ Apache\ Tika\ metadata: ma \\ EMET\ metadata: me \end{array}\right. \\
of\ images\left(\begin{array}{c} FITS\ metadata: mf \\ IPTC\ Photo\ Metadata: mi \\ Metadata\ extractor\ metadata: mm \\ Apache\ Tika\ metadata: ma \\ EMET\ metadata: me \end{array}\right. \\
of\ audio\left(\begin{array}{c} FITS\ metadata: mf \\ Apache\ Tika\ metadata: ma \\ EMET\ metadata: me \end{array}\right. \\
of\ video\left(\begin{array}{c} FITS\ metadata: mf \\ Apache\ Tika\ metadata: ma \\ EMET\ metadata: me \\ Metadata\ extractor\ metadata: mm \end{array}\right.
\end{array}\right.
\end{array}\right.$$

- $\mathcal{D}$ standing for “the class of documents”;
- $\mathcal{S}$ standing for “the class of metadata extraction systems”;
- $\mathcal{M}$ standing for “the class of metadata sets”.

It is also indicated with:

- $\mathcal{D}_t, \mathcal{D}_i,\ \mathcal{D}_a, \mathcal{D}_v$ subclasses of the documents class $\mathcal{D}$;
- $\mathcal{S}_{from\_t}, \mathcal{S}_{from\_i},\ \mathcal{S}_{from\_a}, \mathcal{S}_{from\_v}$ subclasses of the metadata extraction systems $\mathcal{S}$;
- $\mathcal{M}_{of\_t}, \mathcal{M}_{of\_i},\ \mathcal{M}_{of\_a}, \mathcal{M}_{of\_v}$ subclasses of the metadata sets class $\mathcal{M}$

In addition, objects belonging to the document class will instantiate specific properties related to the format $\mathbb{F}$. The format $\mathbb{F}$ denotes the set of formats that can be instantiated by an object x: PDF, DOC, DOCX, PAGES, BMP, GIF, JPEG, MP3, BFW and MP4.

A few terminological remarks are in order: what does "object" mean here? The term will be used as applying to whatever bears properties. An object has properties; by having them it may, as philosophers often say, satisfy certain predicates that denote the properties at issue or, equivalently, make true the corresponding sentences. Documents are objects, for they are property-bearers.

In particular, an object belonging to the subclass $\mathcal{D}_t$ can have a format of type $PDF, DOC, DOCX, PAGES$; to the subclass $\mathcal{D}_i$ can instantiate the formats $BMP, GIF, JPEG, MP3, BFW$ and $FLAC$; to the subclass $\mathcal{D}_a$ can instantiate the format $MP3$ and to the subclass $\mathcal{D}_v$ can instantiate the formats BFW and $MP4$. Given an object x belonging to class $\mathcal{D}$ we write $\mathbb{F}x$ the fact that x instantiates a property of the format $\mathbb{F}$. For example:

- $\mathbb{F}_{PDF}x$ standing for "x is PDF"[71].

Additionally, we have two relations between classes:

- $\mathcal{S}$ ext $\mathcal{M}$ standing for "$\mathcal{S}$ extracts $\mathcal{M}$";
- $\mathcal{S}$ ext_from $\mathcal{D}$ standing for "$\mathcal{S}$ extracts from $\mathcal{D}$".

[71] Of course, "is" is not to be understood as the *is_a* relation representing the links formed in a hierarchical classification of entities.

3.3. The MADME Approach

DMO can be defined as a heavyweight ontology, since it includes classes, subclasses, relazionships between classes, istances, axioms, constraints, theorems and, especially, decision rules[72].

This section will provide a first draft of a formal characterization in first-order logic of the main notions and relations presented in the first Section 3.1. First, let us introduce some axioms about our digital objects. These first axioms serve to establish the belonging of an object to a certain class and the disjunction between classes:

$\mathcal{D}\colon \{x|x \text{ is a document}\}$

$$\text{(A1) } \forall x \left(x \in \mathcal{D} \leftrightarrow (x \notin \mathcal{S} \land x \notin \mathcal{M}) \land \left(x \in \mathcal{D}_t \veebar x \in \mathcal{D}_i \veebar x \in \mathcal{D}_a \veebar x \in \mathcal{D}_v \right) \right)$$

$\mathcal{S}\colon \{x|x \text{ is a metadata extraction system}\}$

$$\text{(A2) } \forall x \left(\begin{matrix} x \in \mathcal{S} \leftrightarrow \\ (x \notin \mathcal{D} \land x \notin \mathcal{M}) \land \left(x \in \mathcal{S}_{from\_t} \veebar x \in \mathcal{S}_{from\_i} \veebar x \in \mathcal{S}_{from\_a} \veebar x \in \mathcal{S}_{from\_v} \right) \end{matrix} \right)$$

$\mathcal{M}\colon \{x|x \text{ is a metadata set}\}$

$$\text{(A3) } \forall x \left(\begin{matrix} x \in \mathcal{M} \leftrightarrow \\ (x \notin \mathcal{D} \land x \notin \mathcal{S}) \land \left(x \in \mathcal{M}_{of\_t} \veebar x \in \mathcal{M}_{of\_i} \veebar x \in \mathcal{M}_{of\_a} \veebar x \in \mathcal{M}_{of\_v} \right) \end{matrix} \right)$$

Axiom (A1) states that an object belongs to class $\mathcal{D}$ if and only if it does not belong to class $\mathcal{S}$ or $\mathcal{M}$ and belongs to subclass $\mathcal{D}_t$ or $\mathcal{D}_i$ or $\mathcal{D}_a$ or $\mathcal{D}_v$. Axiom (A2) states that an object belongs to class $\mathcal{S}$ if and only it does not belong to class $\mathcal{D}$ or $\mathcal{M}$ and belongs to subclass $\mathcal{S}_{from\_t}$ or $\mathcal{S}_{from\_i}$ or $\mathcal{S}_{from\_a}$

[72] An excellent book on the construction of formal ontologies is Arp *et al.* (2015).

or $\mathcal{S}_{from\_v}$. Lastly, axiom (A3) states that an object belongs to class $\mathcal{M}$ if and only if it does not belong to class $\mathcal{D}$ or $\mathcal{S}$ and belongs to subclass $\mathcal{M}_{of\_t}$ or $\mathcal{M}_{of\_i}$ or $\mathcal{M}_{of\_a}$ or $\mathcal{M}_{of\_v}$.

The document class $\mathcal{D}$ will also also have these specific axioms:

(A4) $\mathbb{F}\mathcal{D} \ \rightarrow \exists x(x \in \mathcal{D} \wedge \mathbb{F}x)$

(A5) $\forall x\big(x \in \mathcal{D}_t \rightarrow \mathbb{F}_{PDF}x \veebar \mathbb{F}_{DOC}x \veebar \mathbb{F}_{DOCX}x \veebar \ \mathbb{F}_{PAGES}x \ \big)$

(A6)$\forall x\big(x \in \mathcal{D}_i$
$\rightarrow \mathbb{F}_{BMO}x \veebar \ \mathbb{F}_{GIF}x \veebar \mathbb{F}_{JPEG}x \veebar \ \mathbb{F}_{MP3}x \veebar \mathbb{F}_{BFW}x \veebar \ \mathbb{F}_{FLAC}x\big)$

(A7)$\forall x(x \in \mathcal{D}_a \rightarrow \mathbb{F}_{MP3}x)$

(A8)$\forall x\big(x \in \mathcal{D}_v \rightarrow \mathbb{F}_{BFW}x \veebar \ \mathbb{F}_{MP4}x\big)$

Axiom (A4) states that given $\mathbb{F}\mathcal{D}$ then there is an object x belonging to class $\mathcal{D}$ and x instantiates a property of the format $\mathbb{F}$. Axioms (A5), (A6), (A7) and (A8) on the other hand, determine, based on the subclass to which an object x belongs, which formats $\mathbb{F}$ that object can instantiate.

The theorem (T1) below follows from (A2):

$$(T1)\ \forall x\left(\begin{matrix} x \in \mathcal{S} \\ \rightarrow \big(x = c \veebar x = o \veebar x = g \veebar x = f \veebar x = a \veebar x = e \veebar x = i \veebar x = m\big)\end{matrix}\right)$$

The theorem (T2) below follows from (A3):

$$(T2)\ \forall x\left(\begin{matrix} x \in \mathcal{M} \\ \rightarrow \begin{pmatrix} x = mc \veebar x = mo \veebar x = mg \veebar x = mf \veebar \\ x = ma \veebar x = me \veebar x = mi \veebar x = mm \end{pmatrix}\end{matrix}\right)$$

Some axioms on the relationships between classes are now introduced:

$\mathcal{S}$ ext $\mathcal{M}$ standing for “$\mathcal{S}$ extracts $\mathcal{M}$”:

$(\mathrm{A9})\ \mathcal{S}\ \text{ext}\ \mathcal{M}\ \ \rightarrow \exists x, y(x \in \mathcal{S} \wedge y \in \mathcal{M})$

If there is a metadata extraction system, the metadata set must exist. The relationship is asymmetrical and irreflexive.

$\mathcal{S}$ ext_from $\mathcal{D}$ standing for "$\mathcal{S}$ extracts from $\mathcal{D}$":

$(\mathrm{A10})\ \ \ \mathcal{S}\ \text{ext\_from}\ \mathcal{D} \rightarrow \exists x, y(x \in \mathcal{S} \wedge y \in \mathcal{D})$

$$(\mathrm{A11})\forall x\left(\begin{array}{c} x \in \mathcal{S} \\ \rightarrow \Big(\exists y(y \in \mathcal{D} \wedge x\ \text{ext\_from}\ y)\ \underline{\vee}\ \neg\exists y(y \in \mathcal{D} \wedge x\ \text{ext\_from}\ y)\Big)\end{array}\right)$$

If there is a document object, it may be that a metadata extraction system exists. The relationship is asymmetrical and irreflexive

The theorem follows from $(\mathrm{A9})$ and $(\mathrm{A10})$: $(\mathrm{T3})\ x\ \text{ext\_from}\ y \rightarrow \exists z(z \in \mathcal{M} \wedge x\ \text{ext}\ z\ )$

Based on these axioms and theorems, the following decision rules will be generated:

$(\mathrm{R1})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = c \wedge y\ \text{ext\_from}\ x)\big)$

$(\mathrm{R2})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = o \wedge y\ \text{ext\_from}\ x)\big)$

$(\mathrm{R3})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF} \rightarrow \exists y(y \in \mathcal{S} \wedge y = g \wedge y\ \text{ext\_from}\ x)\big)$

$(\mathrm{R4})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF} \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y\ \text{ext\_from}\ x)\big)$

$(\mathrm{R5})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF} \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y\ \text{ext\_from}\ x\ )\big)$

$(\mathrm{R6})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y\ \text{ext\_from}\ x)\big)$

$(\mathrm{R7})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{DOC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge$
$y\ \text{ext\_from}\ x\ )\big)$

$(\mathrm{R8})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{DOCX}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge$
$y\ \text{ext\_from}\ x\ )\big)$

$$(\mathrm{R9})\forall x\big(x \in \mathcal{D}_t \wedge \mathbb{F}_{PAGES}x \rightarrow \neg\exists y(y \in \mathcal{S} \wedge y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R10})\exists x(x \in \mathcal{S} \wedge x = c \wedge x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mc \wedge x \text{ ext } z\ )$$

$$(\mathrm{R11})\exists x(x \in \mathcal{S} \wedge x = o \wedge\ x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mo \wedge x \text{ ext } z\ \ )$$

$$(\mathrm{R12})\exists x(x \in \mathcal{S} \wedge x = g \wedge x \text{ ext\_from } y\ \ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mg \wedge\ x \text{ ext } z\ )$$

$$(\mathrm{R13})\exists x(x \in \mathcal{S} \wedge x = f \wedge x \text{ ext\_from } y\ \ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mf \wedge\ x \text{ ext } z\ )$$

$$(\mathrm{R14})\exists x(x \in \mathcal{S} \wedge x = a \wedge\ x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = ma \wedge x \text{ ext } z\ \ )$$

$$(\mathrm{R15})\exists x(x \in \mathcal{S} \wedge x = e \wedge x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = me \wedge\ x \text{ ext } z\ )$$

$$(\mathrm{R16})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BMP}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R17})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BMP}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R18})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BMP}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R19})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BMP}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R20})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BMP}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R21})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{GIF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R22})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{GIF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x)\big)$$

$$(\mathrm{R23})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{GIF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R24})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{GIF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R25})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{GIF}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R26})\forall x\left(x \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x\ )\right)$$

$$(\mathrm{R27})\forall x\left(x \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x\ )\right)$$

$$(\mathrm{R28})\forall x\left(x \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x\ )\right)$$

$$(\mathrm{R29})\forall x\left(x \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x\ )\right)$$

$$(\mathrm{R30})\forall x\left(x \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x\ )\right)$$

$$(\mathrm{R31})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R32})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R33})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R34})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R35})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R36})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BFW}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R37})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BFW}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R38})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BFW}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R39})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BFW}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R40})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{BFW}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x\,)\big)$$

$$(\mathrm{R41})\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{FLAC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x\,)\big)$$

(R42)$\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{FLAC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = i \wedge y \text{ ext\_from } x\ )\big)$

(R43)$\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{FLAC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = m \wedge y \text{ ext\_from } x\ )\big)$

(R44)$\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{FLAC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x\ )\big)$

(R45)$\forall x\big(x \in \mathcal{D}_i \wedge \mathbb{F}_{FLAC}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x\ )\big)$

(R46)$\exists x(x \in \mathcal{S} \wedge x = f \wedge x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mf \wedge x \text{ ext } z\ )$

(R47)$\exists x(x \in \mathcal{S} \wedge x = i \wedge\ x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mi \wedge x \text{ ext } z\ \ )$

(R48)$\exists x(x \in \mathcal{S} \wedge x = m \wedge x \text{ ext\_from } y\ \ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mm \wedge\ x \text{ ext } z\ )$

(R49)$\exists x(x \in \mathcal{S} \wedge x = a \wedge\ x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = ma \wedge x \text{ ext } z\ \ )$

(R50)$\exists x(x \in \mathcal{S} \wedge x = e \wedge x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = me \wedge\ x \text{ ext } z\ )$

(R51)$\forall x\big(x \in \mathcal{D}_a \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = f \wedge y \text{ ext\_from } x)\big)$

(R52)$\forall x\big(x \in \mathcal{D}_a \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = a \wedge y \text{ ext\_from } x)\big)$

(R53)$\forall x\big(x \in \mathcal{D}_a \wedge \mathbb{F}_{MP3}x \rightarrow \exists y(y \in \mathcal{S} \wedge y = e \wedge y \text{ ext\_from } x)\big)$

$$(\mathrm{R54})\exists x(x \in \mathcal{S} \land x = f \land x \text{ ext\_from } y) \\ \rightarrow \exists z(z \in \mathcal{M} \land z = mf \land x \text{ ext } z)$$

$$(\mathrm{R55})\exists x(x \in \mathcal{S} \land x = a \land \ x \text{ ext\_from } y) \\ \rightarrow \exists z(z \in \mathcal{M} \land z = ma \land x \text{ ext } z\ )$$

$$(\mathrm{R56})\exists x(x \in \mathcal{S} \land x = e \land x \text{ ext\_from } y\ ) \\ \rightarrow \exists z(z \in \mathcal{M} \land z = me \land \ x \text{ ext } z)$$

$$(\mathrm{R57})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{BFW}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = f \land y \text{ ext\_from } x)\big)$$

$$(\mathrm{R58})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{BFW}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = a \land y \text{ ext\_from } x)\big)$$

$$(\mathrm{R59})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{BFW}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = e \land y \text{ ext\_from } x)\big)$$

$$(\mathrm{R60})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{BFW}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = m \land y \text{ ext\_from } x)\big)$$

$$(\mathrm{R61})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{MP4}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = f \land y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R62})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{MP4}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = a \land y \text{ ext\_from } x)\big)$$

$$(\mathrm{R63})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{MP4}x \rightarrow \exists y(y \in \mathcal{S} \land y = e \land \\ y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R64})\forall x\big(x \in \mathcal{D}_v \land \mathbb{F}_{MP4}x \\ \rightarrow \exists y(y \in \mathcal{S} \land y = m \land y \text{ ext\_from } x\ )\big)$$

$$(\mathrm{R65})\exists x(x \in \mathcal{S} \land x = f \land x \text{ ext\_from } y) \\ \rightarrow \exists z(z \in \mathcal{M} \land z = mf \land x \text{ ext } z)$$

$$(\text{R66})\exists x(x \in \mathcal{S} \wedge x = a \wedge \ x \text{ ext\_from } y\ ) \\ \rightarrow \exists z(z \in \mathcal{M} \wedge z = ma \wedge x \text{ ext } z\ \ )$$

$$(\text{R67})\exists x(x \in \mathcal{S} \wedge x = e \wedge x \text{ ext\_from } y\ \ ) \\ \rightarrow \exists z(z \in \mathcal{M} \wedge z = me \wedge \ x \text{ ext } z\ )$$

$(\text{R68})\exists x(x \in \mathcal{S} \wedge x = m \wedge \ x \text{ ext\_from } y\ ) \rightarrow \exists z(z \in \mathcal{M} \wedge z = mm \wedge x \text{ ext } z\ \ )$

The MADME procedure provides a set of methodological guidelines for the application of DM rules. Decisions have multiple alternatives and there is a need to examine these alternatives in a structured manner. The MADME procedure involves the following steps:

Step 1. Evaluate the type of document source.

i. Identify the class in $\mathcal{D}$.

ii. Identify the format $\mathbb{F}$.

Step 2. Apply decision rules.

Step 3. Evaluate the extraction systems proposed by the procedure.

3.4. MADME in Action

This section will show how the MADME approach makes its choices in practice. Specifically, the examples that will be considered will concern text documents and images from scientific articles on COVID-19 published between 2020 and 2022, while the examples relating to audio and video documents will be taken from the CNR webtv[73] and will always concern COVID-19.

Example 3.4.1. Given a digital document source $\varphi =$ *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[74].

Based on MADME procedure the first step is the evaluation of the type of documentary source. In this case: *i*) $\varphi \in \mathcal{D}_t$ and *ii*) $\mathbb{F}_{PDF}\varphi$. For this reason, $\varphi \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}\varphi$ will be the premise. The second step is the application of decision rules

1.	$\varphi \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}\varphi$	Premise
	$\Downarrow$	
2.	$y_1 \in \mathcal{S} \wedge y_1 = c$	1-R1
	$\Downarrow$	
3.	c ext_from φ	1,2-R1
	$\Downarrow$	
4.	$y_2 \in \mathcal{S} \wedge y_2 = o$	1-R2
	$\Downarrow$	
5.	o ext_from φ	1,4-R2
	$\Downarrow$	
6.	$y_3 \in \mathcal{S} \wedge y_3 = g$	1-R3
	$\Downarrow$	
7.	g ext_from φ	1,6-R3
	$\Downarrow$	
8.	$y_4 \in \mathcal{S} \wedge y_4 = f$	1-R4
	$\Downarrow$	
9.	f ext_from φ	1,8-R4

[73] https://www.cnrweb.tv/.

[74] Hua Li *et al*. (2020).

$\Downarrow$

10. $y_5 \in \mathcal{S} \land y_5 = a$ 1-R5

$\Downarrow$

11. a ext_from φ 1,10-R5

$\Downarrow$

12. $y_6 \in \mathcal{S} \land y_6 = e$ 1-R6

$\Downarrow$

13. e ext_from φ 1,12-R6

$\Downarrow$

14. $z_1 \in \mathcal{M} \land z_1 = mc$ 2,3-R10

$\Downarrow$

15. c ext mc 14-R10

$\Downarrow$

16. $z_2 \in \mathcal{M} \land z_2 = mo$ 4,5-R11

$\Downarrow$

17. o ext mo 16-R11

$\Downarrow$

18. $z_3 \in \mathcal{M} \land z_3 = mg$ 6,7-R12

$\Downarrow$

19. g ext mg 18-R12

$\Downarrow$

20. $z_4 \in \mathcal{M} \land z_4 = mf$ 8,9-R13

$\Downarrow$

21. f ext mf 20-R13

$\Downarrow$

22. $z_5 \in \mathcal{M} \land z_5 = ma$ 10,11-R14

$\Downarrow$

23. a ext ma 22-R14

$\Downarrow$

24. $z_6 \in \mathcal{M} \land z_6 = me$ 12,13-R15

$$\Downarrow$$

25. e ext me 24-R15

Finally, the third step evaluates the metadata extraction systems proposed by the procedure

$$\begin{array}{llllll}
 & \rightarrow & c \text{ ext\_from } \varphi & \rightarrow & c \text{ ext } mc \\
 & \rightarrow & o \text{ ext\_from } \varphi & \rightarrow & o \text{ ext } mo \\
\varphi & \rightarrow & g \text{ ext\_from } \varphi & \rightarrow & g \text{ ext } mg \\
 & \rightarrow & f \text{ ext\_from } \varphi & \rightarrow & f \text{ ext } mf \\
 & \rightarrow & a \text{ ext\_from } \varphi & \rightarrow & a \text{ ext } ma \\
 & \rightarrow & e \text{ ext\_from } \varphi & \rightarrow & e \text{ ext } me
\end{array}$$

The MADME approach allows to establish on the basis of axioms, theorems and rules that if the input document is $\varphi =$ *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[75] the choice will fall on the following metadata extraction systems: CERMINE, OCR++, GROBID, FITS, Apache Tika and EMET.

Example 3.4.2. Given a digital document source $\varphi \in \mathcal{D}_t \land \mathbb{F}_{PAGES}\varphi$:

1. $\varphi \in \mathcal{D}_t \land \mathbb{F}_{PAGES}\varphi$ Premise

$$\Downarrow$$

2. $\otimes$ 1-R9

In the second example the decision procedure is immediately blocked by rule (R9). In this case, the MADME approach states

[75] Hua Li *et al*. (2020).

that the set of possible choices is nothing other than the empty set $\emptyset$.

Example 3.4.3. Given a digital document source $\varphi =$ figure 1 from *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[76].

Based on MADME procedure $\varphi \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}\varphi$ will be the premise.

1.	$\varphi \in \mathcal{D}_i \wedge \mathbb{F}_{JPEG}\varphi$	Premise
	$\Downarrow$	
2.	$y_1 \in \mathcal{S} \wedge y_1 = f$	1-R26
	$\Downarrow$	
3.	f ext_from φ	1,2-R26
	$\Downarrow$	
4.	$y_2 \in \mathcal{S} \wedge y_2 = i$	1-R27
	$\Downarrow$	
5.	i ext_from φ	1,4-R27
	$\Downarrow$	
6.	$y_3 \in \mathcal{S} \wedge y_3 = m$	1-R28
	$\Downarrow$	
7.	m ext_from φ	1,6-R28
	$\Downarrow$	
8.	$y_4 \in \mathcal{S} \wedge y_4 = a$	1-R29
	$\Downarrow$	
9.	a ext_from φ	1,8-R29
	$\Downarrow$	
10.	$y_5 \in \mathcal{S} \wedge y_5 = e$	1-R30

[76] Image extrapolated from *figure 1* from the article Hua Li *et al*. (2020, p. 6560).

$\Downarrow$

11. e ext_from φ 1,10-R30

$\Downarrow$

12. $z_1 \in \mathcal{M} \wedge z_1 = mf$ 2,3-R46

$\Downarrow$

13. f ext mf 12-R46

$\Downarrow$

14. $z_2 \in \mathcal{M} \wedge z_2 = mi$ 4,5-R47

$\Downarrow$

15. i ext mi 14-R47

$\Downarrow$

16. $z_3 \in \mathcal{M} \wedge z_3 = mm$ 6,7-R48

$\Downarrow$

17. m ext mm 16-R48

$\Downarrow$

18. $z_4 \in \mathcal{M} \wedge z_4 = ma$ 8,9-R49

$\Downarrow$

19. a ext ma 18-R13

$\Downarrow$

20. $z_5 \in \mathcal{M} \wedge z_5 = me$ 10,11-R14

$\Downarrow$

21. e ext me 20-R50

The MADME approach allows to establish that if the input document is $\varphi =$ figure 1 from *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[77]the choice will

[77] Image extrapolated from *figure 1* from the article Hua Li *et al*. (2020, p. 6560).

fall on the following extraction systems: FITS, IPTC, Metadata Extractor and Apache Tika

$$\begin{array}{cccccc} & \rightarrow & f \text{ ext\_from } \varphi & \rightarrow & f \text{ ext } mf \\ & \rightarrow & i \text{ ext\_from } \varphi & \rightarrow & i \text{ ext } mi \\ \varphi & \rightarrow & m \text{ ext\_from } \varphi & \rightarrow & m \text{ ext } mm \\ & \rightarrow & a \text{ ext\_from } \varphi & \rightarrow & a \text{ ext } ma \end{array}$$

Example 3.4.4. Given a digital document source $\varphi = (audio\ taken\ from\ "\textit{la pandemia influenza le parole}")$[78].

Based on MADME procedure $\varphi \in \mathcal{D}_a \wedge \mathbb{F}_{MP3}\varphi$ will be the premise

$$\begin{array}{lcl} 1. & \varphi \in \mathcal{D}_a \wedge \mathbb{F}_{MP3}\varphi & \text{Premise} \\ & \Downarrow & \\ 2. & y_1 \in \mathcal{S} \wedge y_1 = f & \text{1-R51} \\ & \Downarrow & \\ 3. & f \text{ ext\_from } \varphi & \text{1,2-R51} \\ & \Downarrow & \\ 4. & y_2 \in \mathcal{S} \wedge y_2 = a & \text{1-R52} \\ & \Downarrow & \\ 5. & a \text{ ext\_from } \varphi & \text{1,4-R52} \\ & \Downarrow & \\ 6. & y_3 \in \mathcal{S} \wedge y_3 = e & \text{1-R53} \\ & \Downarrow & \\ 7. & e \text{ ext\_from } \varphi & \text{1,6-R53} \\ & \Downarrow & \\ 8. & z_1 \in \mathcal{M} \wedge z_1 = mf & \text{2,3-R54} \end{array}$$

[78] https://www.cnrweb.tv/la-pandemia-influenza-le-parole/.

$\Downarrow$

9. $f \text{ ext } mf$ 8-R54

$\Downarrow$

10. $z_2 \in \mathcal{M} \wedge z_2 = ma$ 4,5-R55

$\Downarrow$

11. $a \text{ ext } ma$ 10-R55

$\Downarrow$

12. $z_3 \in \mathcal{M} \wedge z_3 = me$ 6,7-R56

$\Downarrow$

13. $e \text{ ext } me$ 12-R56

The MADME approach allows to establish that if the input document is $\varphi = (audio\ taken\ from$ "*la pandemia influenza le parole"*) the choice will fall on the following extraction systems: FITS, Apache Tika and EMET

$$\begin{tikzcd}
 & f \text{ ext\_from } \varphi \arrow[r] & f \text{ ext } mf \\
\varphi \arrow[ru] \arrow[r] \arrow[rd] & a \text{ ext\_from } \varphi \arrow[r] & a \text{ ext } ma \\
 & e \text{ ext\_from } \varphi \arrow[r] & e \text{ ext } me
\end{tikzcd}$$

Example 3.4.5. Given a digital document source $\varphi = (video\ taken\ from\ "Geografia\ del\ Covid")$[79].

Based on MADME procedure $\varphi \in \mathcal{D}_v \wedge \mathbb{F}_{MP4}\varphi$ will be the premise

1. $\varphi \in \mathcal{D}_v \wedge \mathbb{F}_{MP4}\varphi$ Premise

$\Downarrow$

[79] https://www.cnrweb.tv/geografia-del-covid/.

2. $y_1 \in \mathcal{S} \wedge y_1 = f$ 1-R61

$\Downarrow$

3. f ext_from φ 1,2-R61

$\Downarrow$

4. $y_2 \in \mathcal{S} \wedge y_2 = a$ 1-R62

$\Downarrow$

5. a ext_from φ 1,4-R62

$\Downarrow$

6. $y_3 \in \mathcal{S} \wedge y_3 = e$ 1-R63

$\Downarrow$

7. e ext_from φ 1,6-R63

$\Downarrow$

8. $y_4 \in \mathcal{S} \wedge y_4 = m$ 1-R64

$\Downarrow$

9. m ext_from φ 1,8-R64

$\Downarrow$

10. $z_1 \in \mathcal{M} \wedge z_1 = mf$ 2,3-R65

$\Downarrow$

11. f ext mf 10-R65

$\Downarrow$

12. $z_2 \in \mathcal{M} \wedge z_2 = ma$ 4,5-R66

$\Downarrow$

13. a ext ma 12-R66

$\Downarrow$

14. $z_3 \in \mathcal{M} \wedge z_3 = me$ 6,7-R67

$\Downarrow$

15. e ext me 14-R67

$\Downarrow$

16. $z_4 \in \mathcal{M} \wedge z_4 = mm$ 8,9-R68

$$\Downarrow$$

17. $m \text{ ext } mm$ 16-R68

The MADME procedure allows to that if the input document is $\varphi = (video\ taken\ from\ "\mathit{Geografia\ del\ Covid}")$ the choice will fall on the following extraction systems: FITS, Apache Tika, EMET and Metadata Extractor.

$$\begin{array}{lllll}
 & \rightarrow & f \text{ ext\_from } \varphi & \rightarrow & f \text{ ext } mf \\
\varphi & \rightarrow & a \text{ ext\_from } \varphi & \rightarrow & a \text{ ext } ma \\
 & \rightarrow & e \text{ ext\_from } \varphi & \rightarrow & e \text{ ext } me \\
 & \rightarrow & m \text{ ext\_from } \varphi & \rightarrow & m \text{ ext } mm
\end{array}$$

3.5. An Application Case

Having observed how the procedure works with different document sources, let us take a closer look at example 1. In this example, given the document $\varphi =$ *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[80], the premise was obtained $\varphi \in \mathcal{D}_t \wedge \mathbb{F}_{PDF}\varphi$ from which the following extraction systems can be output: CERMINE, OCR++, GROBID, FITS, Apache Tika and EMET. Now, when the procedure generates multiple choices, it is appropriate to operate according to the following principle:

[80] Hua Li *et al*. (2020).

If given a φ document there is a choice between different extraction systems, then choose, whenever possible, the one with the best metadata extraction percentages.

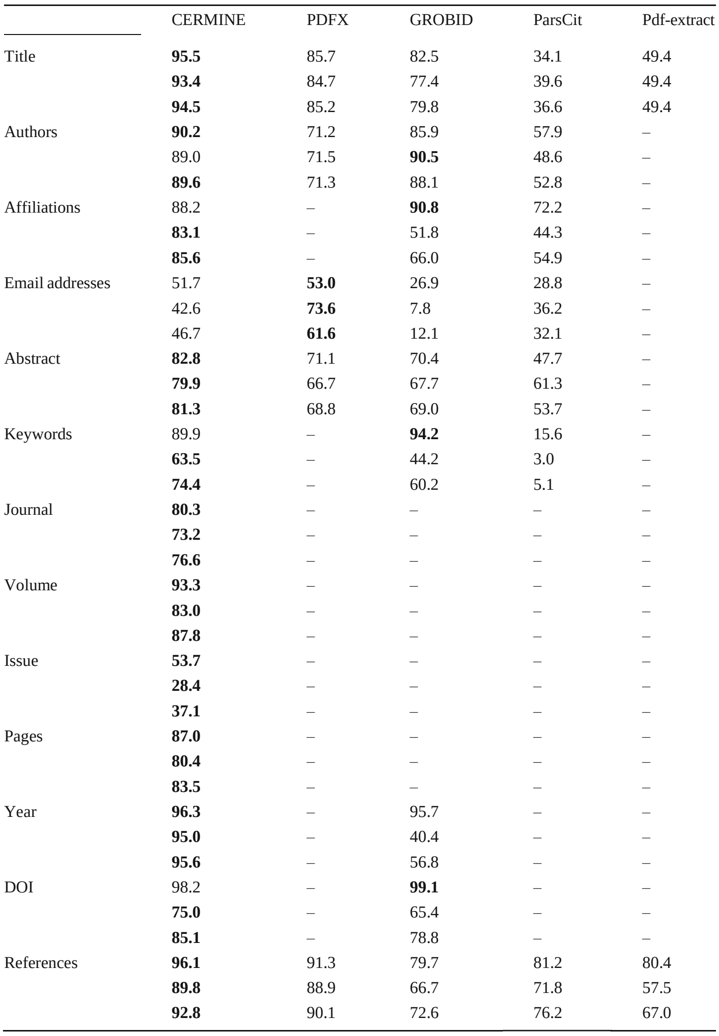
The principle is inspired by the famous Ockham's razor: "*Entia non sunt multiplicanda praeter necessitatem*[81]" or "*Pluralitas non est ponenda sine necessitate*[82]". In the application context of this disserattion, the razor is a powerful problem-solving principle that allows us to optimise and maximise the chances of correctly extracting metadata by choosing the best possible extraction system.

In this case, considering the results of the scientific literature, the principle will opt for CERMINE. Table 3.5.1.[83] compares the best metadata extraction systems given the document φ

[81] Entities must not be multiplied beyond necessity.

[82] Plurality should not be posited without necessity.

[83] Table extrapolated from Tkaczyk *et al*. (2015, p. 333).

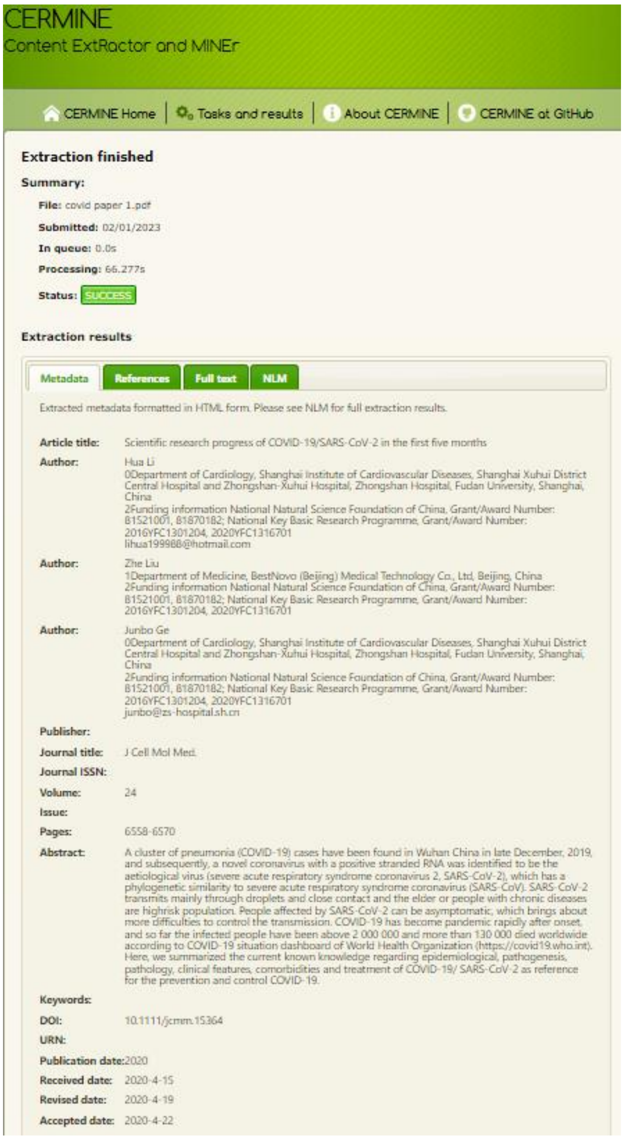


	CERMINE	PDFX	GROBID	ParsCit	Pdf-extract
Title	**95.5**	85.7	82.5	34.1	49.4
	93.4	84.7	77.4	39.6	49.4
	94.5	85.2	79.8	36.6	49.4
Authors	**90.2**	71.2	85.9	57.9	–
	89.0	71.5	**90.5**	48.6	–
	89.6	71.3	88.1	52.8	–
Affiliations	88.2	–	**90.8**	72.2	–
	83.1	–	51.8	44.3	–
	85.6	–	66.0	54.9	–
Email addresses	51.7	**53.0**	26.9	28.8	–
	42.6	**73.6**	7.8	36.2	–
	46.7	**61.6**	12.1	32.1	–
Abstract	**82.8**	71.1	70.4	47.7	–
	79.9	66.7	67.7	61.3	–
	81.3	68.8	69.0	53.7	–
Keywords	89.9	–	**94.2**	15.6	–
	63.5	–	44.2	3.0	–
	74.4	–	60.2	5.1	–
Journal	**80.3**	–	–	–	–
	73.2	–	–	–	–
	76.6	–	–	–	–
Volume	**93.3**	–	–	–	–
	83.0	–	–	–	–
	87.8	–	–	–	–
Issue	**53.7**	–	–	–	–
	28.4	–	–	–	–
	37.1	–	–	–	–
Pages	**87.0**	–	–	–	–
	80.4	–	–	–	–
	83.5	–	–	–	–
Year	**96.3**	–	95.7	–	–
	95.0	–	40.4	–	–
	95.6	–	56.8	–	–
DOI	98.2	–	**99.1**	–	–
	75.0	–	65.4	–	–
	85.1	–	78.8	–	–
References	**96.1**	91.3	79.7	81.2	80.4
	89.8	88.9	66.7	71.8	57.5
	92.8	90.1	72.6	76.2	67.0

Table 3.5.1. The results of comparing the performance of various metadata extraction systems

Therefore, given as input the document $\varphi =$ *Scientific research progress of COVID-19/SARS-CoV-2 in the first five months*[84], the output produced by CERMINE is an XML file in the NLM JATS format:

[84] Hua Li *et al*. (2020).



Extraction finished

Summary:

File: covid paper 1.pdf

Submitted: 02/01/2023

In queue: 0.0s

Processing: 66.277s

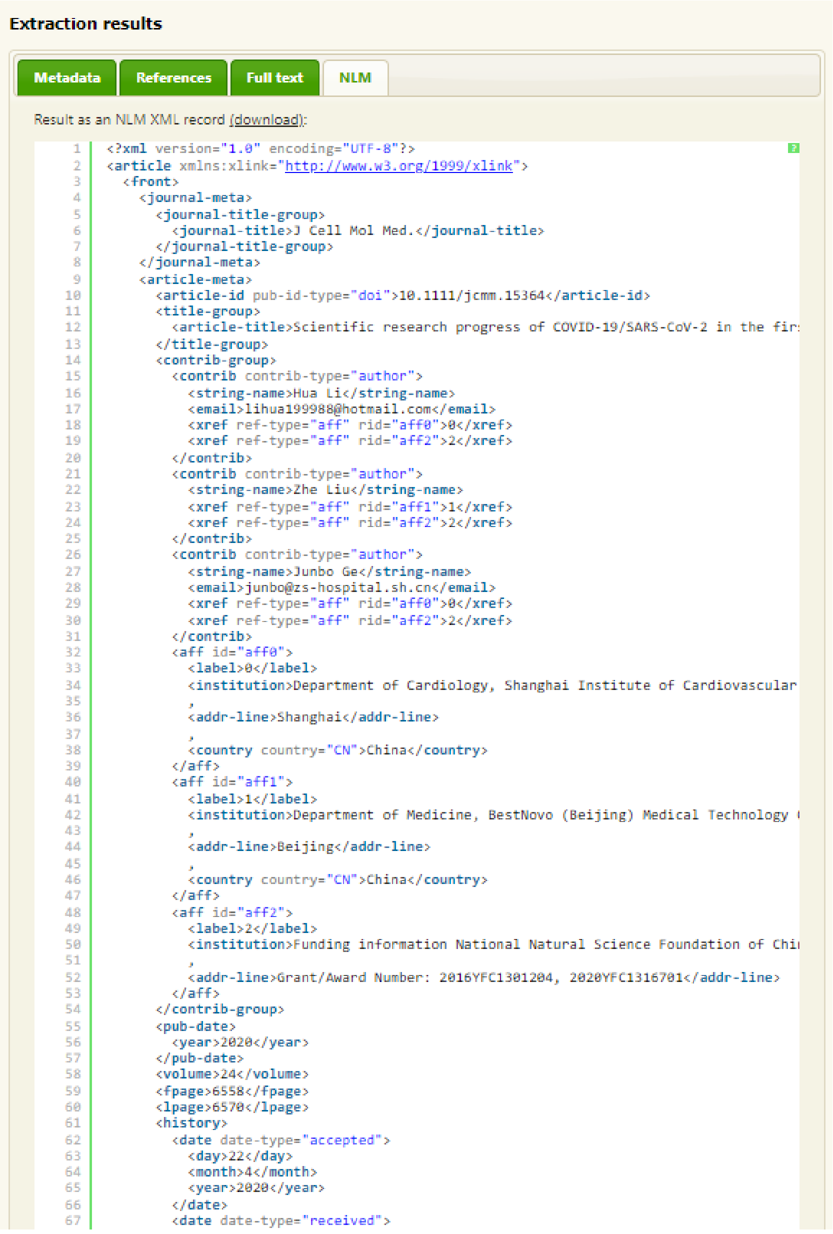
Status: SUCCESS

Extraction results

Extracted metadata formatted in HTML form. Please see NLM for full extraction results.

Article title:	Scientific research progress of COVID-19/SARS-CoV-2 in the first five months
Author:	Hua Li 0Department of Cardiology, Shanghai Institute of Cardiovascular Diseases, Shanghai Xuhui District Central Hospital and Zhongshan-Xuhui Hospital, Zhongshan Hospital, Fudan University, Shanghai, China 2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701 lihua199988@hotmail.com
Author:	Zhe Liu 1Department of Medicine, BestNovo (Beijing) Medical Technology Co., Ltd, Beijing, China 2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701
Author:	Junbo Ge 0Department of Cardiology, Shanghai Institute of Cardiovascular Diseases, Shanghai Xuhui District Central Hospital and Zhongshan-Xuhui Hospital, Zhongshan Hospital, Fudan University, Shanghai, China 2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701 junbo@zs-hospital.sh.cn
Publisher:	
Journal title:	J Cell Mol Med.
Journal ISSN:	
Volume:	24
Issue:	
Pages:	6558-6570
Abstract:	A cluster of pneumonia (COVID-19) cases have been found in Wuhan China in late December, 2019, and subsequently, a novel coronavirus with a positive stranded RNA was identified to be the aetiological virus (severe acute respiratory syndrome coronavirus 2, SARS-CoV-2), which has a phylogenetic similarity to severe acute respiratory syndrome coronavirus (SARS-CoV). SARS-CoV-2 transmits mainly through droplets and close contact and the elder or people with chronic diseases are highrisk population. People affected by SARS-CoV-2 can be asymptomatic, which brings about more difficulties to control the transmission. COVID-19 has become pandemic rapidly after onset, and so far the infected people have been above 2 000 000 and more than 130 000 died worldwide according to COVID-19 situation dashboard of World Health Organization (https://covid19.who.int). Here, we summarized the current known knowledge regarding epidemiological, pathogenesis, pathology, clinical features, comorbidities and treatment of COVID-19/ SARS-CoV-2 as reference for the prevention and control COVID-19.
Keywords:	
DOI:	10.1111/jcmm.15364
URN:	
Publication date:	2020
Received date:	2020-4-15
Revised date:	2020-4-19
Accepted date:	2020-4-22

Figure 3.5.1. Metadata extracted by CERMINE



Result as an NLM XML record (download):

```
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <front>
    <journal-meta>
      <journal-title-group>
        <journal-title>J Cell Mol Med.</journal-title>
      </journal-title-group>
    </journal-meta>
    <article-meta>
      <article-id pub-id-type="doi">10.1111/jcmm.15364</article-id>
      <title-group>
        <article-title>Scientific research progress of COVID-19/SARS-CoV-2 in the fir
      </title-group>
      <contrib-group>
        <contrib contrib-type="author">
          <string-name>Hua Li</string-name>
          <email>lihua199988@hotmail.com</email>
          <xref ref-type="aff" rid="aff0">0</xref>
          <xref ref-type="aff" rid="aff2">2</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Zhe Liu</string-name>
          <xref ref-type="aff" rid="aff1">1</xref>
          <xref ref-type="aff" rid="aff2">2</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Junbo Ge</string-name>
          <email>junbo@zs-hospital.sh.cn</email>
          <xref ref-type="aff" rid="aff0">0</xref>
          <xref ref-type="aff" rid="aff2">2</xref>
        </contrib>
        <aff id="aff0">
          <label>0</label>
          <institution>Department of Cardiology, Shanghai Institute of Cardiovascular
          ,
          <addr-line>Shanghai</addr-line>
          ,
          <country country="CN">China</country>
        </aff>
        <aff id="aff1">
          <label>1</label>
          <institution>Department of Medicine, BestNovo (Beijing) Medical Technology
          ,
          <addr-line>Beijing</addr-line>
          ,
          <country country="CN">China</country>
        </aff>
        <aff id="aff2">
          <label>2</label>
          <institution>Funding information National Natural Science Foundation of Chi
          ,
          <addr-line>Grant/Award Number: 2016YFC1301204, 2020YFC1316701</addr-line>
        </aff>
      </contrib-group>
      <pub-date>
        <year>2020</year>
      </pub-date>
      <volume>24</volume>
      <fpage>6558</fpage>
      <lpage>6570</lpage>
      <history>
        <date date-type="accepted">
          <day>22</day>
          <month>4</month>
          <year>2020</year>
        </date>
        <date date-type="received">
```

Figure 3.5.2. Partial XML of metadata extracted by CERMINE[85]

In this way, the framework extracts, from document φ, mostly Dublin Core metadata (title, author, affiliation, abstract, keywords, journal name, etc.). The next chapter will show how the extracted metadata can be logically modelled.

[85] http://cermine.ceon.pl/cermine/task.html;jsessionid=B6A03EF510C304E2D9A62AE98F5CE729?task=8357847398154435111.

Chapter 4

Epistemic Logic for Metadata Modelling

4.1. The Art of Modelling

More than 10 years have passed since Chris Anderson published an article entitled “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete[86]”. Anderson’s article has quickly become the ideological manifesto of *datacentric enthusiasm* and is articulated along two key points.

First: *trust me, it’s convenient*. Search engines have taught that it is not important to understand why one web page is “better” than another, but it is sufficient to trust the ordering produced by the PageRank algorithm. The convenience of receiving a simple answer to a potentially complicated question, without necessarily having to develop any semantic or causal analysis, has quickly become the key to success for search engines such as Google.

Second: *scientific models are obsolete*. The unprecedented availability of data allows us to rethink radically the relationship between data and the mechanisms generating them. According to Anderson, it is possible to stop looking for models: instead of proceeding by “conjectures and refutations” to explaining observations, the deluge of data allows us to dispense the laborious task of *constructing models* for the phenomena of

[86] Anderson (2008). At that time, Anderson was the chief editor of the influent technology magazine *Wired*.

interest, in favour of the much easier task of analysing the *correlations* identified by sophisticated statistical algorithms.

This dissertation, in agreement with Hosni[87] will move in the opposite direction to that outlined by Anderson.

First: *don't trust*. Today more than ever it is necessary to emphasise the quality of information and the veracity of data: a semantic analysis is necessary.

Second: *scientific models are fundamental*. It is very difficult to think data without them responding to a modelling hypothesis. The simplistic idea that petabytes of data can be self-sufficient and that data can be seen as a substitute for scientific modelling is not sustainable[88].

In particular, in this chapter: *i*) a model based on epistemic logic will be proposed to formalise metadata extracted from scientific articles on COVID-19 by means of automatic extraction systems; *ii*) the issue of data quality will be emphasised through the definition of a *metaontological principle of veracity as truthmaker*. Whereas until a few years ago the cost of information was the most important aspect, today the quality of information has become more important than ever. For this reason, the *veracity* of the information was proposed as the fourth "V" (the other ones being Volume, Variety, and Velocity) of big data[89].

[87] Hosni (2018).

[88] Furthermore, Anderson seems to have no knowledge whatsoever about the fact that the role of models and modelling in scientific research has been exhaustively and rigorously studied by philosophers of science. In particular, on the fact that models can be used to *understand* and *explain* the world see Giere (2004); Bokulich (2011); Weisberg (2013).

[89] M. G. Lozano *et al*. (2020); Lukoianova *et al.* (2014); Snow (2012).

Based on the principle expressed at the end of the last chapter, the automatic metadata extraction systems we be used in this chapter is still CERMINE. In the context of research articles, metadata are usually descriptive in nature and hold great importance as they provide a brief overview of a scientific article by showing, as seen in Chapter 1, information such as its title, authors, journal, bibliography, etc. Often, researchers tend to decide on the relevance of the article to their domain of interest based on the information in the metadata. For this reason, the question of metadata veracity is now a central issue in the world of information.

The next section will show how epistemic logic can be used to model structured metadata and how the tools of metaontology are able to propose a definition of veracity as truthmaker.

4.2. The "Metadata Extraction Logic" Model

As shown in Chapter 2, epistemic logic is an extension of classical logic that has as its object of study the statements of belief and knowledge. Since Hintikka's epistemic logic has been a subject of research in philosophy, computer science, artificial intelligence and game theory. Hintikka provided a semantic interpretation of epistemic and belief operators that can be presented 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.

Assuming a minimal definition of information as "data + semantics", trust on extracted information can be identified with the result of a consistency assessment. In this context, an extracted information is consistent when it allows preserving: i) the set of beliefs and knowledge base of the extraction agent; and ii) the informational properties of the object from which the extraction was performed. Therefore, having to deal with sets of beliefs and knowledge, epistemic logic turns out to be the most suitable logic for this task. In particular, standard epistemic logic can be applied to metadata modelling in the following way[90]. At the syntactic level will be used 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 4.2.1. [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 the language $\mathcal{L}_{K_{\mathcal{E}}}$ will be defined 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 the concept of possible world will be replaced with that of *possible extraction*.

Definition 4.2.2. [Epistemic Model] Given a set $\mathcal{P}_{\mathcal{E}}$ of primitive propositions and a set $\mathcal{F}$ of 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

[90] Cf. Cuconato (2021a, 2021b, 2022).

$R_a \subseteq E \times E$ 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 4.2.3. [Semantics of $\mathcal{L}_{K_\mathcal{E}}$]: Given a model $M{:}\langle E, R^{\mathcal{F}}, V^{\mathcal{P}_\mathcal{E}} \rangle$, a formula φ to be true in (M, e_i), written $M, e_i \vDash \varphi$, will be inductively defined 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 4.2.4. [Axioms and Inference Rules] The proof system of metadata extraction logic model that will be used is axiomatized using the axiom of **T** and the rule of modus ponens and necessitation. The system is presented in Table 4.2.1.

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	$\underbrace{\overset{\overset{a}{\curvearrowright}}{\mathcal{E}^{d_i}_{m_i}}}_{e_i}$

Table 4.2.1. System **T**

Definition 4.2.5. [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).

In more detail, it is possible to systematically determine the truth value of a formula in the structure $\mathcal{S}$. As already known, the truth of a propositional formula depends on "the situation of the world", or in the case of an epistemic proposition "is true in w on condition that it is true in all worlds accessible from w". Situations are formalised using evaluations, and in $\mathcal{S}$ we know that a proposition $\mathcal{P}_\mathcal{E}$ "is true in e on condition that it is true in all possible extractions accessible from e"

$$V^{\mathcal{P}_\mathcal{E}}: E \to (\mathcal{P}_\mathcal{E} \to \{true, false\})$$

Also, since $p_\mathcal{E}$ has the form $\mathcal{E}_{m_i}^{d_i}$ it will be written 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

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

For example, the graph shows 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.

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

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

But what does it mean that in an extraction an extracted metadata is true? Put another way, what does it mean that a framework correctly extracts a metadata? In order to answer these questions, it is necessary to present the *theory of truthmaker* and define veracity as truthmaker. Truthmaker theory is an interesting meta-ontological theory from the world of analytical philosophy that explores the relationships between what is *true* and what *exists.*

The theory has deep roots in Western philosophy and, on the one hand, conveys an emerging intuition of ours: if, for example, it is true that the cat is on the roof it is because the cat is "in fact" on the roof; on the other hand, it represents the idea behind a famous theory of truth, namely *correspondence*[91]. The theory can be summarised by Dummett's[92] regulative principle C (for "Correspondence") is that:

($\mathcal{C}$) If a statement is true, there must be something in virtue of which it is true.

The thing in question is called a truthmaker[93]. As Rodriguez-Pereyra specifies:

> To believe in truthmaking is, basically, to believe that truth is grounded in the world or reality[94].

Put another way, whenever something is true, there must be something which makes it true. Will be specified later what is meant in metadata domain by "something". The theory of truthmaker to be considered in this dissertation is the one developed by the Australian philosopher David Malet Armstrong in *Truth and Truthmakers*[95]:

($\mathcal{T}$) For every truth, p, there exists an entity, T, such that T makes p true if and only if it is not possible that T exists and p is false.

In metadata domain, it is possible to reformulate the Armstrongian $\mathcal{T}$ principle as follows:

[91] A classic example of correspondence theory is the statement by the scholastic philosopher Thomas Aquinas: "*Veritas est adaequatio rei et intellectus*". However, the first occurrence of a basic truthmaking idea is found in Aristotle's *Categories*.

[92] Dummett (1976, p. 89).

[93] Truthmaking has become one of the most important metaphysical topics of the late 20th-century and early 21st-century philosophy. For an introduction to truthmaker theory see Rodriguez-Pereyra (2002, 2005), while for an application of the truthmakers to the modal basis of scientific modelling see Tahko (2023).

[94] Rodriguez-Pereyra (2006, p. 186).

[95] Armstrong (2004). See also Calemi (2014); Cuconato (2014a, 2014b).

$(\mathcal{V})$ For every true proposition, $p_{\mathcal{E}}$, there exists a document, d, such that d makes $p_{\mathcal{E}}$ true if and only if it is not possible for d to exist and $p_{\mathcal{E}}$ to be false.

In this way, it will be possible to fully appreciate the meaning of this principle. Let us again consider our scheme:

$$(1)\,\underbrace{\mathcal{E}_{m_i}^{d_i}}_{e_i} = \mathrm{T/F}$$

Therefore, the typical truthmaking question can be asked: by virtue of what (1) is true? Well, by virtue of $\mathcal{V}$ saying that a framework a has correctly extracted a metadata m means that there is a document d that "makes true" the extraction e.

4.3. Application of Standard Metadata Modelling

This section will consider specific metadata extractions. These first two extractions will focus on four specific metadata – title, author, journal, and publication date.

The first document d_1[96] describes the effectiveness of a second booster vaccine against hospitalization and death from COVID-19 in adults aged over 60 years, while the second document d_2[97] explores collective and personal psychiatric trauma related to COVID-19.

Consider the following structure $\mathcal{S}_1 = \langle \mathcal{F}, E, \mathcal{P}_{\mathcal{E}}, M, D \rangle$:

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

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

[96] Arbel *et al*. (2022).
[97] Kalsched (2021).

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

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

$D = \{d_1, d_2\}$

Given document d_1 and MEA a (CERMINE) the following scenario occurs:

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

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

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

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

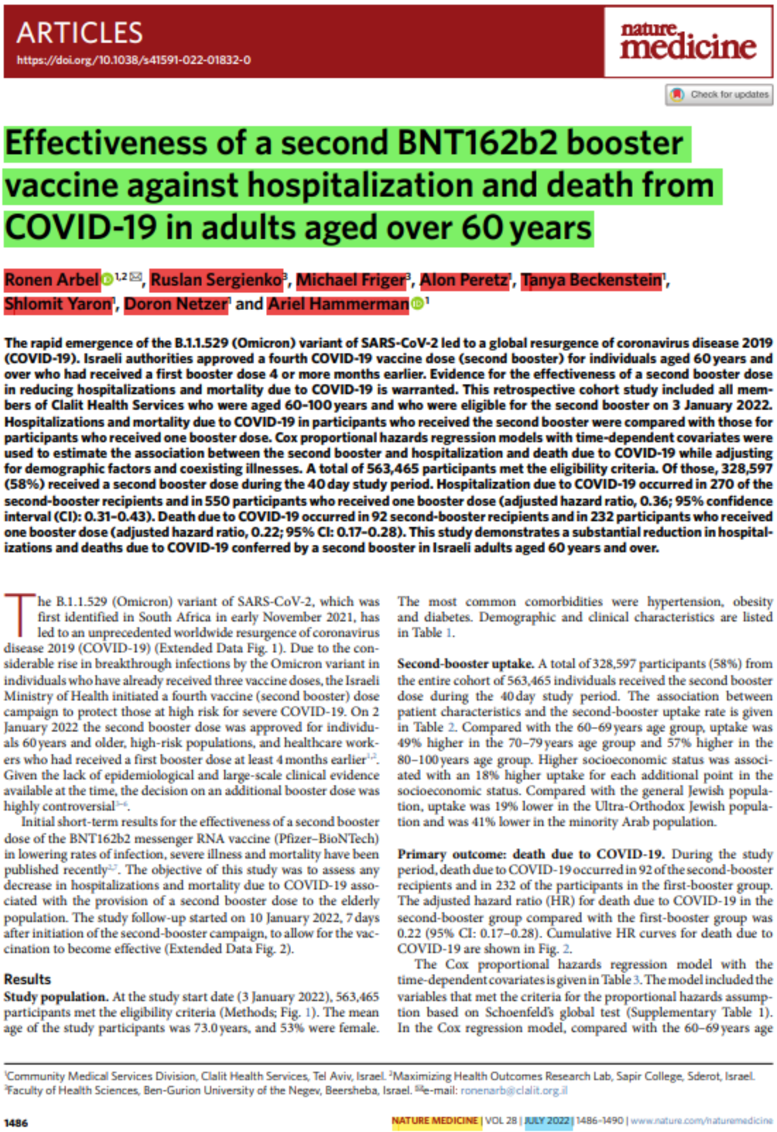
With the first document d_1 extraction e_4 occurs[98]:

- $\underbrace{\mathcal{E}_{m_1}^{d_1}}_{e_4} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_2}^{d_1}}_{e_4} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_3}^{d_1}}_{e_4} = \mathrm{F}$

[98] http://cermine.ceon.pl/cermine/task.html;jsessionid=5B7ECF36991887536E51E9034227C91E?task=5936518302167758199.

- $\underbrace{\mathcal{E}_{m_4}^{d_1}}_{e_4} = \mathrm{T}$

It is now necessary to analyse this extraction in detail. In Figure 4.3.1., the metadata “title” is highlighted in green, the “author” metadata in red, the metadata “journal” in yellow and the metadata “publication date” in blue, while Figures 4.3.2. and 4.3.3. show the extraction metadata results formatted in HTML form and as an NLM XML record.



ARTICLES

https://doi.org/10.1038/s41591-022-01832-0

nature medicine

Check for updates

Effectiveness of a second BNT162b2 booster vaccine against hospitalization and death from COVID-19 in adults aged over 60 years

Ronen Arbel[1,2] ✉, Ruslan Sergienko[3], Michael Friger[3], Alon Peretz[1], Tanya Beckenstein[1], Shlomit Yaron[1], Doron Netzer[1] and Ariel Hammerman[1]

The rapid emergence of the B.1.1.529 (Omicron) variant of SARS-CoV-2 led to a global resurgence of coronavirus disease 2019 (COVID-19). Israeli authorities approved a fourth COVID-19 vaccine dose (second booster) for individuals aged 60 years and over who had received a first booster dose 4 or more months earlier. Evidence for the effectiveness of a second booster dose in reducing hospitalizations and mortality due to COVID-19 is warranted. This retrospective cohort study included all members of Clalit Health Services who were aged 60–100 years and who were eligible for the second booster on 3 January 2022. Hospitalizations and mortality due to COVID-19 in participants who received the second booster were compared with those for participants who received one booster dose. Cox proportional hazards regression models with time-dependent covariates were used to estimate the association between the second booster and hospitalization and death due to COVID-19 while adjusting for demographic factors and coexisting illnesses. A total of 563,465 participants met the eligibility criteria. Of those, 328,597 (58%) received a second booster dose during the 40 day study period. Hospitalization due to COVID-19 occurred in 270 of the second-booster recipients and in 550 participants who received one booster dose (adjusted hazard ratio, 0.36; 95% confidence interval (CI): 0.31–0.43). Death due to COVID-19 occurred in 92 second-booster recipients and in 232 participants who received one booster dose (adjusted hazard ratio, 0.22; 95% CI: 0.17–0.28). This study demonstrates a substantial reduction in hospitalizations and deaths due to COVID-19 conferred by a second booster in Israeli adults aged 60 years and over.

The B.1.1.529 (Omicron) variant of SARS-CoV-2, which was first identified in South Africa in early November 2021, has led to an unprecedented worldwide resurgence of coronavirus disease 2019 (COVID-19) (Extended Data Fig. 1). Due to the considerable rise in breakthrough infections by the Omicron variant in individuals who have already received three vaccine doses, the Israeli Ministry of Health initiated a fourth vaccine (second booster) dose campaign to protect those at high risk for severe COVID-19. On 2 January 2022 the second booster dose was approved for individuals 60 years and older, high-risk populations, and healthcare workers who had received a first booster dose at least 4 months earlier[1,2]. Given the lack of epidemiological and large-scale clinical evidence available at the time, the decision on an additional booster dose was highly controversial[3–6].

Initial short-term results for the effectiveness of a second booster dose of the BNT162b2 messenger RNA vaccine (Pfizer–BioNTech) in lowering rates of infection, severe illness and mortality have been published recently[2,7]. The objective of this study was to assess any decrease in hospitalizations and mortality due to COVID-19 associated with the provision of a second booster dose to the elderly population. The study follow-up started on 10 January 2022, 7 days after initiation of the second-booster campaign, to allow for the vaccination to become effective (Extended Data Fig. 2).

Results

Study population. At the study start date (3 January 2022), 563,465 participants met the eligibility criteria (Methods; Fig. 1). The mean age of the study participants was 73.0 years, and 53% were female. The most common comorbidities were hypertension, obesity and diabetes. Demographic and clinical characteristics are listed in Table 1.

Second-booster uptake. A total of 328,597 participants (58%) from the entire cohort of 563,465 individuals received the second booster dose during the 40 day study period. The association between patient characteristics and the second-booster uptake rate is given in Table 2. Compared with the 60–69 years age group, uptake was 49% higher in the 70–79 years age group and 57% higher in the 80–100 years age group. Higher socioeconomic status was associated with an 18% higher uptake for each additional point in the socioeconomic status. Compared with the general Jewish population, uptake was 19% lower in the Ultra-Orthodox Jewish population and was 41% lower in the minority Arab population.

Primary outcome: death due to COVID-19. During the study period, death due to COVID-19 occurred in 92 of the second-booster recipients and in 232 of the participants in the first-booster group. The adjusted hazard ratio (HR) for death due to COVID-19 in the second-booster group compared with the first-booster group was 0.22 (95% CI: 0.17–0.28). Cumulative HR curves for death due to COVID-19 are shown in Fig. 2.

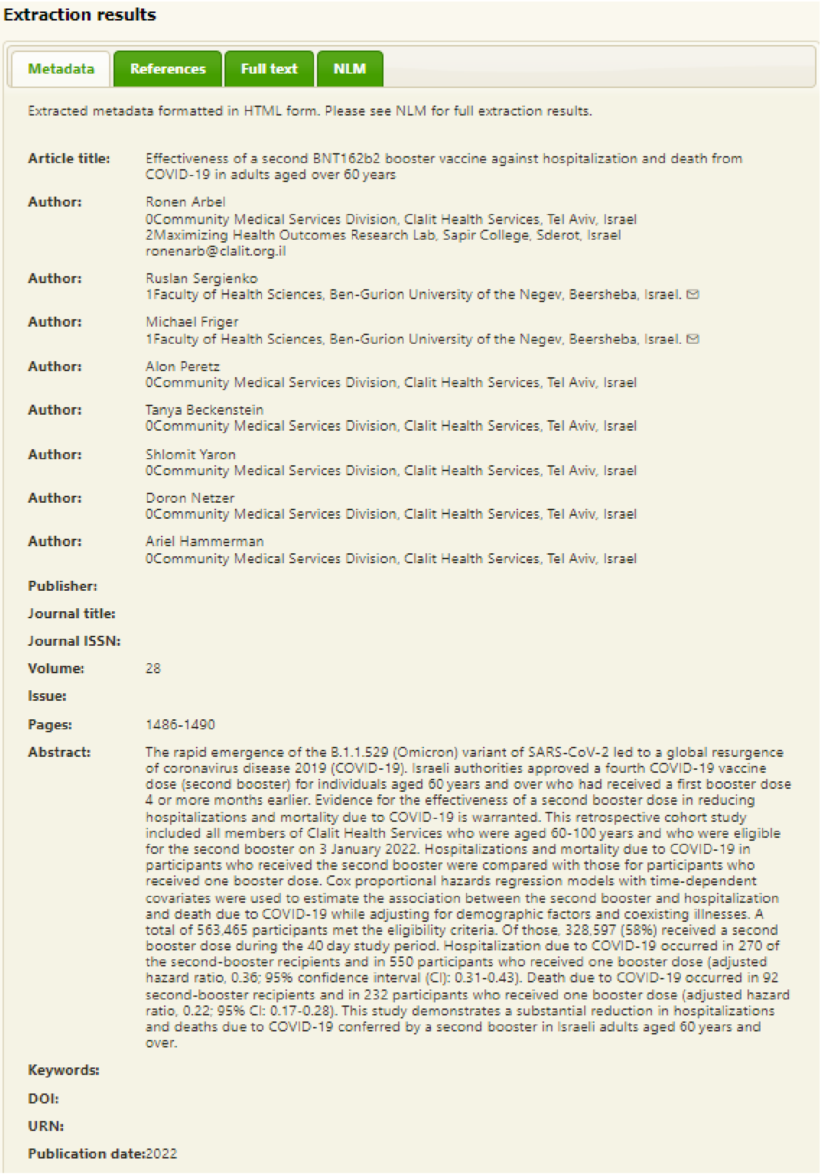
The Cox proportional hazards regression model with the time-dependent covariates is given in Table 3. The model included the variables that met the criteria for the proportional hazards assumption based on Schoenfeld's global test (Supplementary Table 1). In the Cox regression model, compared with the 60–69 years age

[1]Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel. [2]Maximizing Health Outcomes Research Lab, Sapir College, Sderot, Israel. [3]Faculty of Health Sciences, Ben-Gurion University of the Negev, Beersheba, Israel. ✉e-mail: ronenarb@clalit.org.il

1486

NATURE MEDICINE | VOL 28 | JULY 2022 | 1486–1490 | www.nature.com/naturemedicine

Figure 4.3.1. d_1 with highlighted metadata

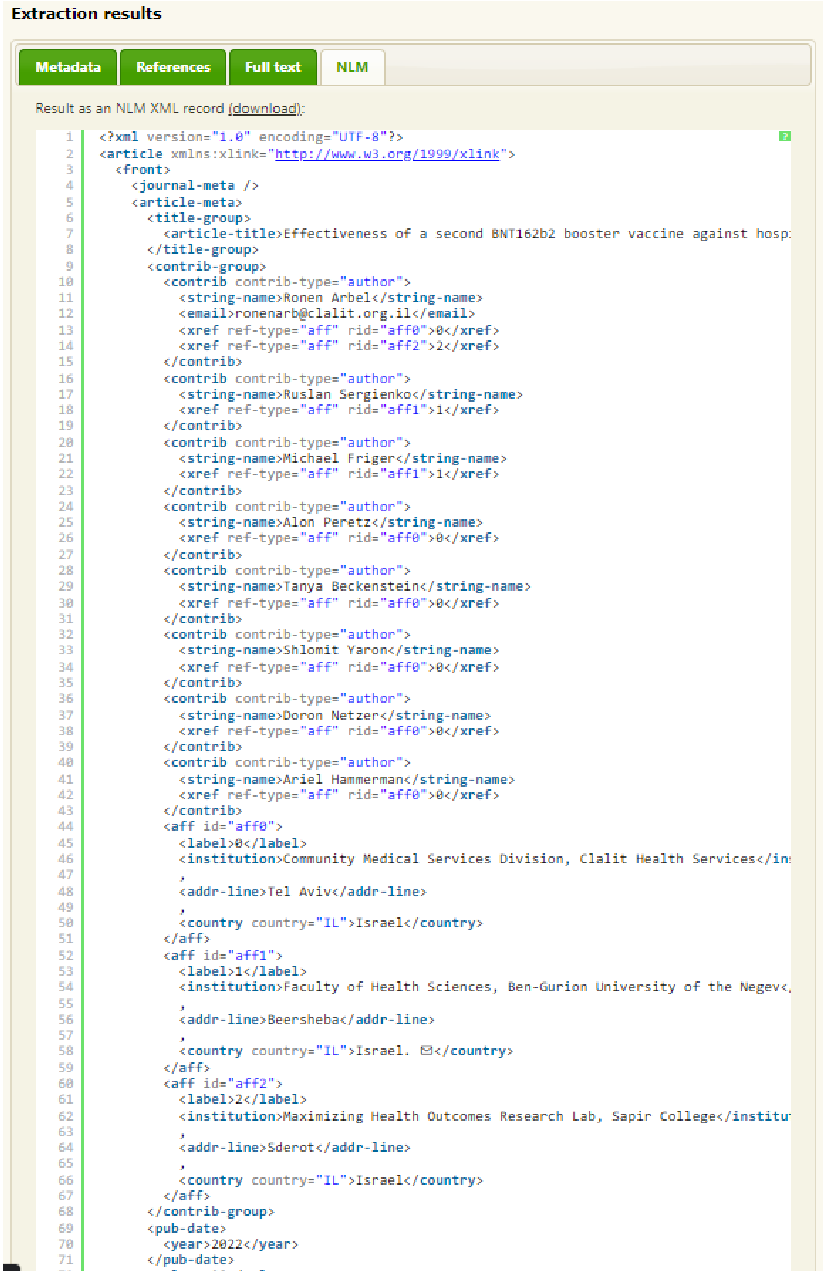


Extraction results

Extracted metadata formatted in HTML form. Please see NLM for full extraction results.

Article title:	Effectiveness of a second BNT162b2 booster vaccine against hospitalization and death from COVID-19 in adults aged over 60 years
Author:	Ronen Arbel 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel 2Maximizing Health Outcomes Research Lab, Sapir College, Sderot, Israel ronenarb@clalit.org.il
Author:	Ruslan Sergienko 1Faculty of Health Sciences, Ben-Gurion University of the Negev, Beersheba, Israel. ✉
Author:	Michael Friger 1Faculty of Health Sciences, Ben-Gurion University of the Negev, Beersheba, Israel. ✉
Author:	Alon Peretz 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
Author:	Tanya Beckenstein 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
Author:	Shlomit Yaron 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
Author:	Doron Netzer 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
Author:	Ariel Hammerman 0Community Medical Services Division, Clalit Health Services, Tel Aviv, Israel
Publisher:	
Journal title:	
Journal ISSN:	
Volume:	28
Issue:	
Pages:	1486-1490
Abstract:	The rapid emergence of the B.1.1.529 (Omicron) variant of SARS-CoV-2 led to a global resurgence of coronavirus disease 2019 (COVID-19). Israeli authorities approved a fourth COVID-19 vaccine dose (second booster) for individuals aged 60 years and over who had received a first booster dose 4 or more months earlier. Evidence for the effectiveness of a second booster dose in reducing hospitalizations and mortality due to COVID-19 is warranted. This retrospective cohort study included all members of Clalit Health Services who were aged 60-100 years and who were eligible for the second booster on 3 January 2022. Hospitalizations and mortality due to COVID-19 in participants who received the second booster were compared with those for participants who received one booster dose. Cox proportional hazards regression models with time-dependent covariates were used to estimate the association between the second booster and hospitalization and death due to COVID-19 while adjusting for demographic factors and coexisting illnesses. A total of 563,465 participants met the eligibility criteria. Of those, 328,597 (58%) received a second booster dose during the 40 day study period. Hospitalization due to COVID-19 occurred in 270 of the second-booster recipients and in 550 participants who received one booster dose (adjusted hazard ratio, 0.36; 95% confidence interval (CI): 0.31-0.43). Death due to COVID-19 occurred in 92 second-booster recipients and in 232 participants who received one booster dose (adjusted hazard ratio, 0.22; 95% CI: 0.17-0.28). This study demonstrates a substantial reduction in hospitalizations and deaths due to COVID-19 conferred by a second booster in Israeli adults aged 60 years and over.
Keywords:	
DOI:	
URN:	
Publication date:	2022

Figure 4.3.2. Metadata extracted by CERMINE from d_1



Result as an NLM XML record (download):

```
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <front>
    <journal-meta />
    <article-meta>
      <title-group>
        <article-title>Effectiveness of a second BNT162b2 booster vaccine against hosp
      </title-group>
      <contrib-group>
        <contrib contrib-type="author">
          <string-name>Ronen Arbel</string-name>
          <email>ronenarb@clalit.org.il</email>
          <xref ref-type="aff" rid="aff0">0</xref>
          <xref ref-type="aff" rid="aff2">2</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Ruslan Sergienko</string-name>
          <xref ref-type="aff" rid="aff1">1</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Michael Friger</string-name>
          <xref ref-type="aff" rid="aff1">1</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Alon Peretz</string-name>
          <xref ref-type="aff" rid="aff0">0</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Tanya Beckenstein</string-name>
          <xref ref-type="aff" rid="aff0">0</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Shlomit Yaron</string-name>
          <xref ref-type="aff" rid="aff0">0</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Doron Netzer</string-name>
          <xref ref-type="aff" rid="aff0">0</xref>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Ariel Hammerman</string-name>
          <xref ref-type="aff" rid="aff0">0</xref>
        </contrib>
        <aff id="aff0">
          <label>0</label>
          <institution>Community Medical Services Division, Clalit Health Services</in
          ,
          <addr-line>Tel Aviv</addr-line>
          ,
          <country country="IL">Israel</country>
        </aff>
        <aff id="aff1">
          <label>1</label>
          <institution>Faculty of Health Sciences, Ben-Gurion University of the Negev<,
          ,
          <addr-line>Beersheba</addr-line>
          ,
          <country country="IL">Israel. ✉</country>
        </aff>
        <aff id="aff2">
          <label>2</label>
          <institution>Maximizing Health Outcomes Research Lab, Sapir College</institu
          ,
          <addr-line>Sderot</addr-line>
          ,
          <country country="IL">Israel</country>
        </aff>
      </contrib-group>
      <pub-date>
        <year>2022</year>
      </pub-date>
```

Figure 4.3.3. Partial XML of metadata extracted by CERMINE from d_1

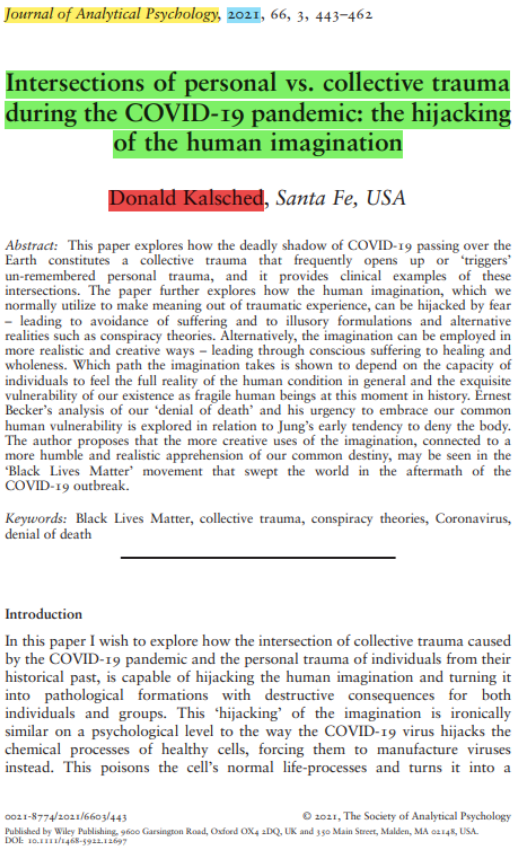
With the second document d_2 extraction e_3 occurs[99]:

- $\underbrace{\mathcal{E}_{m_1}^{d_2}}_{e_3} = \mathrm{T}$

[99] http://cermine.ceon.pl/cermine/task.html?task=5288760465930537804.

- $\underbrace{\mathcal{E}_{m_2}^{d_2}}_{e_3} = \mathrm{F}$
- $\underbrace{\mathcal{E}_{m_3}^{d_2}}_{e_3} = \mathrm{T}$
- $\underbrace{\mathcal{E}_{m_4}^{d_2}}_{e_3} = \mathrm{T}$

In Figure 4.3.4., the metadata "title" is highlighted in green, the "author" metadata in red, the metadata "journal" in yellow and the metadata "publication date" in blue, while Figures 4.3.5. and 4.3.6. show the extraction metadata results formatted in HTML form and as an NLM XML record.



Journal of Analytical Psychology, 2021, 66, 3, 443–462

Intersections of personal vs. collective trauma during the COVID-19 pandemic: the hijacking of the human imagination

Donald Kalsched, *Santa Fe, USA*

Abstract: This paper explores how the deadly shadow of COVID-19 passing over the Earth constitutes a collective trauma that frequently opens up or 'triggers' un-remembered personal trauma, and it provides clinical examples of these intersections. The paper further explores how the human imagination, which we normally utilize to make meaning out of traumatic experience, can be hijacked by fear – leading to avoidance of suffering and to illusory formulations and alternative realities such as conspiracy theories. Alternatively, the imagination can be employed in more realistic and creative ways – leading through conscious suffering to healing and wholeness. Which path the imagination takes is shown to depend on the capacity of individuals to feel the full reality of the human condition in general and the exquisite vulnerability of our existence as fragile human beings at this moment in history. Ernest Becker's analysis of our 'denial of death' and his urgency to embrace our common human vulnerability is explored in relation to Jung's early tendency to deny the body. The author proposes that the more creative uses of the imagination, connected to a more humble and realistic apprehension of our common destiny, may be seen in the 'Black Lives Matter' movement that swept the world in the aftermath of the COVID-19 outbreak.

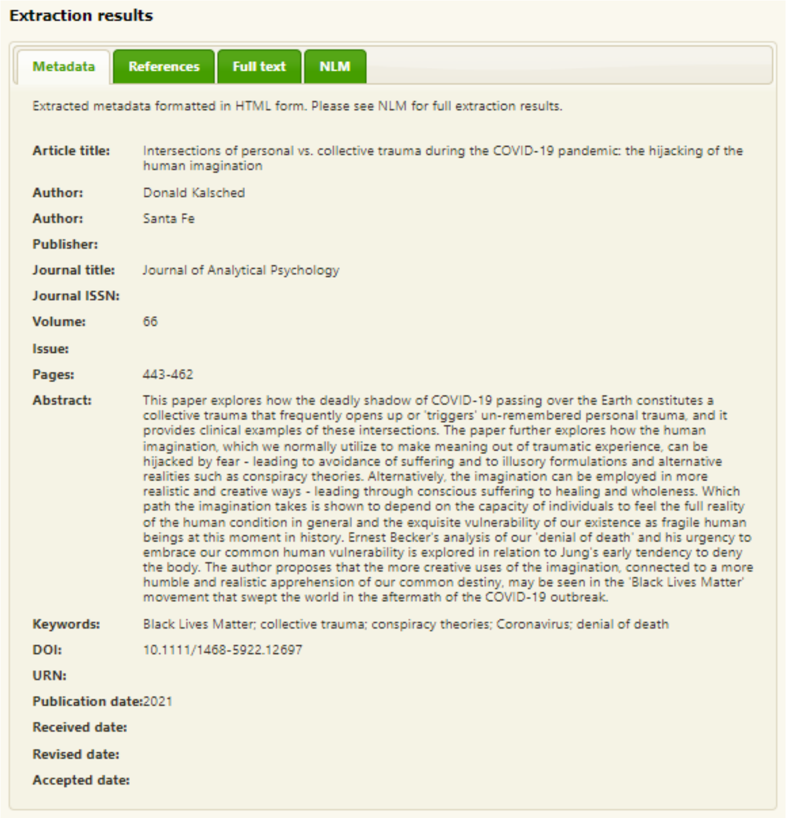
Keywords: Black Lives Matter, collective trauma, conspiracy theories, Coronavirus, denial of death

Introduction

In this paper I wish to explore how the intersection of collective trauma caused by the COVID-19 pandemic and the personal trauma of individuals from their historical past, is capable of hijacking the human imagination and turning it into pathological formations with destructive consequences for both individuals and groups. This 'hijacking' of the imagination is ironically similar on a psychological level to the way the COVID-19 virus hijacks the chemical processes of healthy cells, forcing them to manufacture viruses instead. This poisons the cell's normal life-processes and turns it into a

0021-8774/2021/6603/443 © 2021, The Society of Analytical Psychology
Published by Wiley Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.
DOI: 10.1111/1468-5922.12697

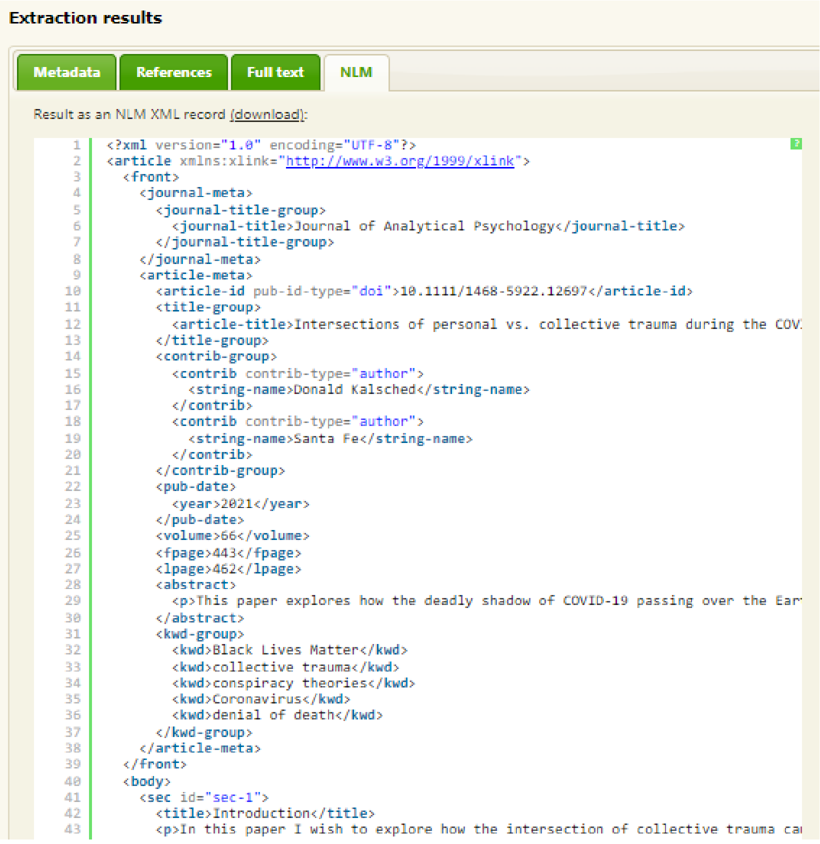
Figure 4.3.4. d_2 with highlighted metadata



Extracted metadata formatted in HTML form. Please see NLM for full extraction results.

Article title:	Intersections of personal vs. collective trauma during the COVID-19 pandemic: the hijacking of the human imagination
Author:	Donald Kalsched
Author:	Santa Fe
Publisher:	
Journal title:	Journal of Analytical Psychology
Journal ISSN:	
Volume:	66
Issue:	
Pages:	443-462
Abstract:	This paper explores how the deadly shadow of COVID-19 passing over the Earth constitutes a collective trauma that frequently opens up or 'triggers' un-remembered personal trauma, and it provides clinical examples of these intersections. The paper further explores how the human imagination, which we normally utilize to make meaning out of traumatic experience, can be hijacked by fear - leading to avoidance of suffering and to illusory formulations and alternative realities such as conspiracy theories. Alternatively, the imagination can be employed in more realistic and creative ways - leading through conscious suffering to healing and wholeness. Which path the imagination takes is shown to depend on the capacity of individuals to feel the full reality of the human condition in general and the exquisite vulnerability of our existence as fragile human beings at this moment in history. Ernest Becker's analysis of our 'denial of death' and his urgency to embrace our common human vulnerability is explored in relation to Jung's early tendency to deny the body. The author proposes that the more creative uses of the imagination, connected to a more humble and realistic apprehension of our common destiny, may be seen in the 'Black Lives Matter' movement that swept the world in the aftermath of the COVID-19 outbreak.
Keywords:	Black Lives Matter; collective trauma; conspiracy theories; Coronavirus; denial of death
DOI:	10.1111/1468-5922.12697
URN:	
Publication date:	2021
Received date:	
Revised date:	
Accepted date:	

Figure 4.3.5. Metadata extracted by CERMINE from d_2



Extraction results

Result as an NLM XML record (download):

```xml
<?xml version="1.0" encoding="UTF-8"?>
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <front>
    <journal-meta>
      <journal-title-group>
        <journal-title>Journal of Analytical Psychology</journal-title>
      </journal-title-group>
    </journal-meta>
    <article-meta>
      <article-id pub-id-type="doi">10.1111/1468-5922.12697</article-id>
      <title-group>
        <article-title>Intersections of personal vs. collective trauma during the COV
      </title-group>
      <contrib-group>
        <contrib contrib-type="author">
          <string-name>Donald Kalsched</string-name>
        </contrib>
        <contrib contrib-type="author">
          <string-name>Santa Fe</string-name>
        </contrib>
      </contrib-group>
      <pub-date>
        <year>2021</year>
      </pub-date>
      <volume>66</volume>
      <fpage>443</fpage>
      <lpage>462</lpage>
      <abstract>
        <p>This paper explores how the deadly shadow of COVID-19 passing over the Ear
      </abstract>
      <kwd-group>
        <kwd>Black Lives Matter</kwd>
        <kwd>collective trauma</kwd>
        <kwd>conspiracy theories</kwd>
        <kwd>Coronavirus</kwd>
        <kwd>denial of death</kwd>
      </kwd-group>
    </article-meta>
  </front>
  <body>
    <sec id="sec-1">
      <title>Introduction</title>
      <p>In this paper I wish to explore how the intersection of collective trauma ca
```

Figure 4.3.6. Partial XML of metadata extracted by CERMINE from d_2

These metadata extractions can be represented by the model of Figure 4.3.7.

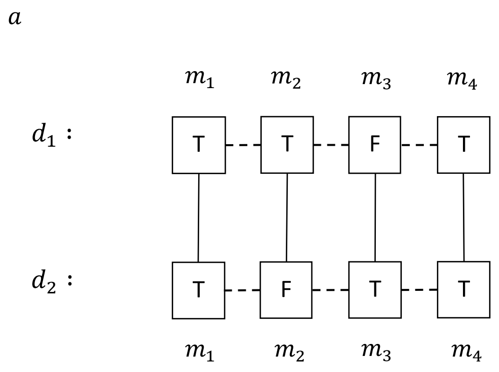


Figure 4.3.7. The model of S_1

4.4. A Four-Valued Epistemic Logic

The logic used in the previous sections provides a strict formal basis and a precise definition of what it means for a metadata to be correctly extracted. However, this approach turns out to be extremely rigid and constrained to only true and false values. Epistemic logic is usually employed to model two aspects of a situation: the factual and the epistemic. The truth, however, is not always attainable and in many cases, it is necessary to reason with partial, incomplete or even incoherent information. This is especially the case in the information sciences, which have to deal with an extremely varied quantity and quality of data. For these reasons, this section will describe a four-valued epistemic logic designed to deal with these situations.

The advantage of using a four-valued epistemic logic is that it does not leave out of the discussion an important factor in the formation of beliefs: *evidence*. Belnap[100] first, and later Dunn[101] and Priest[102], provided an initial interpretation of a four-valued logic, centered precisely on the idea of evidence. In that logic, a proposition p can be, besides true or false, *both* (true and false) or *neither* (true nor false).

Specifically, this section will apply a simplified version of the four-value epistemic logic (FVEL, for short) developed by Santos[103], to the modelling of metadata[104].

[100] Belnap (1977).
[101] Dunn (1976).
[102] Priest (2008).
[103] Santos (2020).
[104] Cf. Cuconato (2023).

Definition 4.4.1. [Syntax of $\mathcal{L}_{FV}$] Let $\mathcal{P}$ be a countable set of atomic propositions and $\mathcal{A}$ a finite set of agents. A well-formed formula φ in our language $\mathcal{L}_{FV}$ is inductively defined as follows:

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

with $p_{\mathcal{E}} \in \mathcal{P}$ and $a \in \mathcal{A}$.

Definition 4.4.2. [Four-Value Epistemic Model] Given a set $\mathcal{P}_{\mathcal{E}}$ of primitive propositions and a set $\mathcal{F}$ of MEA, a four-value epistemic model is a structure $M_{FV} : \langle E, R_{FV}^{\mathcal{F}}, V_{FV}^{\mathcal{P}_{\mathcal{E}}} \rangle$ where

- $E \neq \emptyset$ is a set of possible extractions;
- $R_{FV}^{\mathcal{F}} = (R_{FV1}^{\mathcal{F}}, R_{FV2}^{\mathcal{F}}, \dots, R_{FVn}^{\mathcal{F}})$ is an n-tuple of binary relations on E[105];
- $V_{FV}^{\mathcal{P}_{\mathcal{E}}} : \mathcal{P}_{\mathcal{E}} \times E \to 2^{\{0,1\}}$: is a valuation function that, assigns to each proposition one of four truth values: $\{0\}$ is *false* (f), $\{1\}$ is *true* (t), $\emptyset$ is none (n) and $\{0,1\}$ is *both* (b).

Figure 4.4.1. compares a standard epistemic model with a non-standard four-valued epistemic model, where $\{1\}, \{0\}, \{0,1\}$ and $\emptyset$ mean, in order, true, false, both and none.

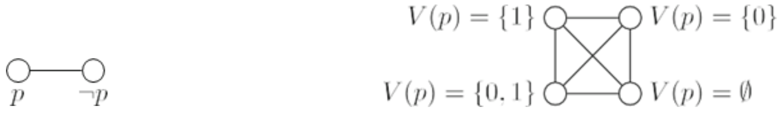


Figure 4.4.1. A standard epistemic model (left) and a non-standard FVEL (right)[106]

[105] In addition, arbitrary accessibility relations will be considered in the remainder of the dissertation.

[106] Figure taken from Santos (2020).

Definition 4.4.3. [Semantics of Syntax of $\mathcal{L}_{FV}$]: With $p_{\mathcal{E}} \in \mathcal{P}$, $e \in E$, $a \in \mathcal{A}$ and $\varphi, \psi \in \mathcal{L}_{FV}$, the satisfaction relation $\vDash$ is inductively defined as follows:

$M_{FV}, e \vDash p_{\mathcal{E}}$	iff	$1 \in V_{FV}^{\mathcal{P}_{\mathcal{E}}}(p_{\mathcal{E}}, e)$
$M_{FV}, e \vDash p_{\mathcal{E}}$	iff	$0 \in V_{FV}^{\mathcal{P}_{\mathcal{E}}}(p_{\mathcal{E}}, e)$
$M_{FV}, e \vDash \varphi \wedge \psi$	iff	$M_{FV}, e \vDash \varphi$ and $M_{FV}, e \vDash \psi$
$M_{FV}, e \vDash \neg(\varphi \wedge \psi)$	iff	$M_{FV}, e \vDash \neg\varphi \; or \; M_{FV}, e \vDash \neg\psi$
$M_{FV}, e \vDash {\sim}\varphi$	iff	$M_{FV}, e \nvDash \varphi$
$M_{FV}, e \vDash \neg{\sim}\varphi$	iff	$M_{FV}, e \vDash \varphi$
$M_{FV}, e \vDash \neg\neg\varphi$	iff	$M_{FV}, e \vDash \varphi$
$M_{FV}, e_1 \vDash K_a\varphi$	iff	for all $e_2 \in E$ $s$uch that $e_1 R e_2$, it holds that $M_{FV}, e_1 \vDash \varphi$

Since the interpretation of formulas is based on the concept of evidence, it is necessary to make some clarifications at the semantic level. Firstly, non-epistemic formulas φ and $\neg\varphi$ are read as *there is evidence for* φ and *there is evidence agains*t φ, respectively. Secondly, the negation $\sim$ is classical: $\sim \varphi$ means that *it is not the case that* φ. Thirdly, the K operator cannot be read in the standard way as *in all possible worlds compatible with what* a *knows, it is the case that* φ , but rather as:

$K_a\varphi$: *agent a knows that there is evidence for* φ.

In this way, it is possible to speak of four-valued formulas in general and define the *extended evaluation function* $\bar{V}_{FV}^{\mathcal{P}_{\mathcal{E}}}: \mathcal{L}_{FV} \times E \to 2^{\{0,1\}}$ as follows:

$$1 \in \bar{V}_{FV}^{\mathcal{P}_{\mathcal{E}}}(\varphi, e) \text{ iff } M_{FV}, e \vDash \varphi$$

$$0 \in \bar{V}_{FV}^{\mathcal{P}_{\mathcal{E}}}(\varphi, e) \text{ iff } M_{FV}, e \vDash \neg\varphi$$

Since the semantics of FVEL is non-compositional, the readings of its formulas will be non-compositional as well. Truth and falsity of formulas are evaluated independently, and for that reason the semantic conditions for each negated formula are defined separately. However, even if the semantics of ¬ is defined on a case-by-case, the connective is still truth-functional[107].

Therefore, one must think of the (four-valued) valuation function as representing evidence or information, while the accessibility relations account for the uncertainty of the agents about which evidential state is the correct one.

Definition 4.4.4. [Four-Valued 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

[107] Santos (2020, p. 457).

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 (in this case, papers on Covid 19).

Compared to definition 4.2.5, since $p_{\mathcal{E}}$ we will denote by $\{1\}, \{0\}, \{0,1\}$ and $\emptyset$, the fact that the information is, respectively, *true*, *false*, *both* and *none*.

$$\mathcal{E}_{m_i}^{d_i} = \{1\}/\{0\}/\{0,1\}/\emptyset$$

4.5. Application of Non-Standard Metadata Modelling

This section will show how metadata modelling changes by applying FVEL. The extraction will be from a scientific article in COVID-19 used at the end of the Chapter 3. The document d_3[108] concerns a medical article presenting the progress of scientific knowledge in the first five months after the start of the pandemic.

Consider the following structure $\mathcal{S}_2 = \langle \mathcal{F}, E, \mathcal{P}_{\mathcal{E}}, M, D \rangle$:

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

$E = \{e_1\}$;

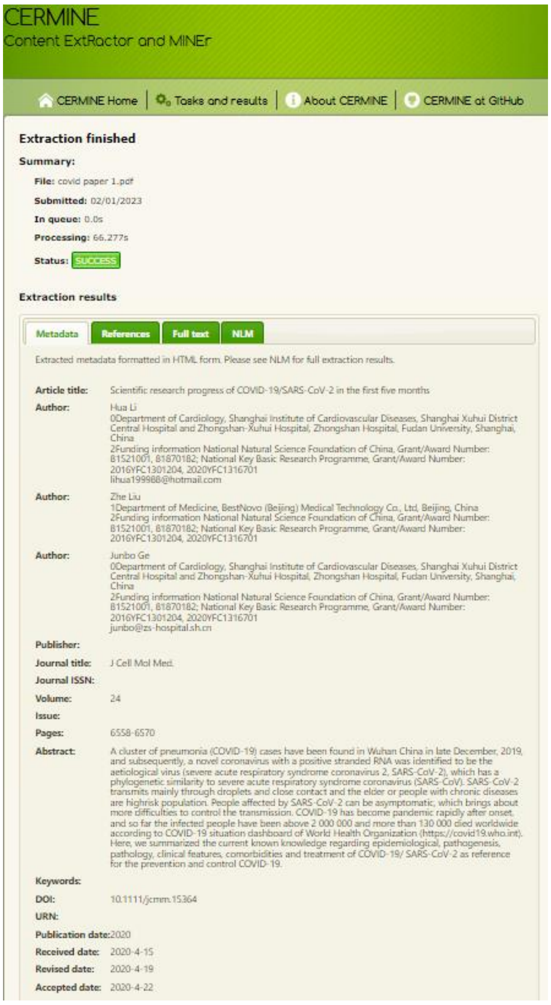
$\mathcal{P}_{\mathcal{E}} = \{p_{\mathcal{E}_1}, \dots, p_{\mathcal{E}_m}\}$

$M = \{m_1, \dots, m_{16}\}$

$D = \{d_3\}$

With the document d_3 MEA a extracts the following metadata:

[108] Hua Li *et al*. (2020).



Extraction finished

Summary:

File: covid paper 1.pdf
Submitted: 02/01/2023
In queue: 0.0s
Processing: 66.277s
Status: SUCCESS

Extraction results

Extracted metadata formatted in HTML form. Please see NLM for full extraction results.

Article title: Scientific research progress of COVID-19/SARS-CoV-2 in the first five months

Author: Hua Li
0Department of Cardiology, Shanghai Institute of Cardiovascular Diseases, Shanghai Xuhui District Central Hospital and Zhongshan-Xuhui Hospital, Zhongshan Hospital, Fudan University, Shanghai, China
2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701
lihua199988@hotmail.com

Author: Zhe Liu
1Department of Medicine, BestNovo (Beijing) Medical Technology Co., Ltd, Beijing, China
2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701

Author: Junbo Ge
0Department of Cardiology, Shanghai Institute of Cardiovascular Diseases, Shanghai Xuhui District Central Hospital and Zhongshan-Xuhui Hospital, Zhongshan Hospital, Fudan University, Shanghai, China
2Funding information National Natural Science Foundation of China, Grant/Award Number: 81521001, 81870182; National Key Basic Research Programme, Grant/Award Number: 2016YFC1301204, 2020YFC1316701
junbo@zs-hospital.sh.cn

Publisher:

Journal title: J Cell Mol Med.

Journal ISSN:

Volume: 24

Issue:

Pages: 6558-6570

Abstract: A cluster of pneumonia (COVID-19) cases have been found in Wuhan China in late December, 2019, and subsequently, a novel coronavirus with a positive stranded RNA was identified to be the aetiological virus (severe acute respiratory syndrome coronavirus 2, SARS-CoV-2), which has a phylogenetic similarity to severe acute respiratory syndrome coronavirus (SARS-CoV). SARS-CoV-2 transmits mainly through droplets and close contact and the elder or people with chronic diseases are highrisk population. People affected by SARS-CoV-2 can be asymptomatic, which brings about more difficulties to control the transmission. COVID-19 has become pandemic rapidly after onset, and so far the infected people have been above 2 000 000 and more than 130 000 died worldwide according to COVID-19 situation dashboard of World Health Organization (https://covid19.who.int). Here, we summarized the current known knowledge regarding epidemiological, pathogenesis, pathology, clinical features, comorbidities and treatment of COVID-19/ SARS-CoV-2 as reference for the prevention and control COVID-19.

Keywords:

DOI: 10.1111/jcmm.15364

URN:

Publication date:2020

Received date: 2020-4-15

Revised date: 2020-4-19

Accepted date: 2020-4-22

Figure 4.5.1. Metadata extracted by CERMINE from d_3

In particular, about the “Author” metadata (m_2), the extracted information is partially correct because, on the one hand, it is true that the author's name is reported correctly, but on the other hand, additional information is reported that is not part of the author's name (such as *affiliation*). Compared to what happened with d_2, this situation can easily be handled within FVEL. In fact, in this specific case:

$$\mathcal{E}_{m_2}^{d_3} = \{0,1\}$$

In this way, a FVEL-based model makes it possible to retain part of the extracted information without necessarily having to consider the extraction of the “Author” metadata completely wrong. Not only that, FVEL is able to classify more accurately even when metadata is completely absent, as in the case of metadata “keywords” (m_{10}). In this case it will be:

$$\mathcal{E}_{m_{10}}^{d_3} = \emptyset$$

This modelling therefore makes it possible to accurately preserve and classify the extracted information even when the extracted metadata is not totally correct.

Conclusions and Future Work

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, the dissertation has moved within an interdisciplinary and multidisciplinary perspective in the fields of logic, knowledge engineering, applied ontology and library and information science. In this way, on the one hand the research had the advantage of drawing on theoretical and technological aspects belonging to different scientific fields, on the other hand it contributed to the dialogue between different and apparently distant scientific sectors. In detail, the following points were developed in the dissertation: *i*) the creation of an innovative logical and ontological framework to develop a decision-making procedure to guide the choice of metadata extraction systems; *ii*) the formal modelling of the extracted metadata; *iii*) the application of classical and non-classical logic to knowledge engineering and library and information science; and *iv*) the application of the framework to specific case studies.

These points clearly bring out the *theoretical* nature of the framework. Theoretical, because the framework is logically and ontologically grounded and presents aspects, techniques, rules and procedures at both syntactic and semantic levels that are absent from the current scientific literature on metadata. However, the theoretical sphere, on the one hand, lay solid and rigorous logical foundations, on the other hand, they highlight the limitations of the more "practical" part concerning the automatic implementation of decision-making processes. For this reason, a future line of research would then be to adapt and implement the present theoretical framework. The presence of a rigorous formal apparatus can facilitate the translation of the procedural rules into a programming language and, consequently, the fully automated development of MADME.

Another line of research could be the extension of metadata verification and modelling to other types of document sources, with the possibility of using and comparing different metadata extraction systems.

Last but not least, the possibility of experimenting with new modal logics to be applied to engineering and library and information science, and addressing difficult questions about the modal basis of scientific modelling, where the central issues concern the nature and justification of the modal content of statements made on the basis of models.

Bibliography

Abdel-Karim, B.M., Pfeuffer, N., Hinz, O. (2021), Machine learning in information systems – a bibliographic review and open research issues, *Electronic Markets*, 31, pp. 643–670. https://doi.org/10.1007/s12525-021-00459-2

Anderson, C. (2008), The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*. http://www.wired.com/2008/06/pb-theory/.

Aristotle, (1923), *Aristotle's Metaphysics*. Greek text and commentary by W.D Ross, vols. 1 and 2. Oxford: Clarendon Press.

Aristotle, (1963), *Aristotle's Categories and De Interpretatione*. Translation and commentary by J.L. Ackrill. Oxford: Clarendon Press.

Aristotle, (1975), *Aristotle's Posterior Analytics*. Translation with notes by J. Barnes. Oxford: Clarendon Press.

Armstrong, D.M. (2004), *Truth and Truthmakers*. Cambridge: Cambridge University Press.

Arp, R., Smith, B., Spear, A.D. (2015), *Building Ontologies with Basic Formal Ontology*. Cambridge, Massachusetts, London, England: The MIT Press. https://doi.org/10.7551/mitpress/9780262527811.001.0001.

Aydin, M.N. (2006), *Decision-making support for method adaptation*. Ed. Enschede. University of Twente, Netherlands.

Belnap, N. (1977), A useful four-valued logic. In J.M. Dunn, G. Epstein (Eds.), *Modern uses of multiple-valued logic*, pp. 5–37. Berlin: Springer.

Berto, F., Plebani, M. (2015), *Ontology and Metaontology*. *A Contemporary Guide*. Bloomsbury.

Bokulich, A. (2011), How Scientific Models Can Explain, *Synthese*,180, n. 1, pp. 33–45. https://doi.org/10.1007/s11229-009-9565-1

Boole, G. (1847), *The Mathematical Analysis of Logic, Being an Essay Towards a Calculus of Deductive Reasoning*. Originally published in Cambridge by Macmillan, B, 1847. Reprinted in Oxford by Basil Blackwell, 1951.

Breitman, K. K., Casanova, M. A., Truszkowski, W. (2007), *Semantic Web: Concepts technologies and applications*. Springer.

Burnett, K., Kwong, B.N., Park, S. (1999), A comparison of the two traditions for metadata development, *Journal of the American Society for Information Science*, 50:13, pp. 1209–1217.

Calegari, R., Ciatto, G., Denti, E., Omicini, A. (2020), Logic-Based Technologies for Intelligent Systems: State of the Art and Perspectives, *Information*, 11, 3: 167. https://doi.org/10.3390/info11030167.

Calemi, F.F. (2013), *Le radici dell'essere. Metafisica e metaontologia in David Malet Armstrong*. Roma: Armando Editore.

Caplan, P. (2003), *Metadata fundamentals for all librarians*. Chicago: American Library Association.

Chan, L.M. (1994), *Cataloging and classification: an introduction*, 2nd ed New York: McGraw-Hill, Inc.

Cuconato, S. (2014a), Mondi di Wittgenstein. Metaontologia del ‘Tractatus’ e teoria dei ‘truthmakers’ di Armstrong. *Rivista Italiana di Filosofia Analitica junior*, 5:2, Special Issue: Metaphysics, pp. 53–65.

Cuconato, S. (2014b), review of “Le Radici dell’essere. Metafisica e metaontologia in D.M. Armstrong” (F.F. Calemi), *Philosophical News*, 9–Humanitas.

Cuconato, S. (2021a), 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.

Cuconato, S. (2021b), Epistemic logic for metadata modelling from scientific papers on COVID-19, *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 9:2, pp. 83–96.

Cuconato, S. (2022a), *Impegno ontologico e l’argomento di indispensabilità in filosofia della matematica. Un’analisi*. Il Sileno Edizioni.

Cuconato, S. (2022b), A logical-metaontological approach to the problem of (meta)data veracity in systems for automatic extraction of metadata from scientific-legal articles, *Science & Philosophy – Journal of Epistemology, Science and Philosophy*, 10:2, pp. 168–187.

Cuconato, S. (2023), A Four-Valued Epistemic Logic for Metadata Modelling from Medical Articles on Pain Therapies, *5th International Conference on Computational Intelligence in Pattern Recognition, Special Session: Intelligent Approaches for Data Mining Applications* (forthcoming).

Cui, B., Chen, X. (2010), An Improved Hidden Markov Model for Literature Metadata Extraction. In: *Advanced Intelligent Computing Theories and Applications,* 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18–21, 2010. Proceedings. 2010, pp. 205–212.

Davenport, T.H., Prosak, L. (2000), *Working knowledge how organisations manage what they know*. Boston: Harvard Busmess School Press.

Dummett, M. (1976), What is a Theory of Meaning? In Gareth Evans and John McDowell, editors, *Truth and Meaning*, pp. 67–137, Oxford University Press.

Dunn, J. (1976), Intuitive semantics for first-degree entailments and 'coupled trees', *Philosophical Studies,* 29:3, pp. 149–168.

Eden, B.L. (2002), *Metadata and its application*. Chicago: ALA TechSource.

Floridi, L. (2009), Against digital ontology. *Synthese*, 168:1, pp. 51–178.

Frege, G. (1879), *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle a. S.: Louis Nebert. Translated as Concept Script, a formal language of pure thought modelled upon that of arithmetic, by S. Bauer-Mengelberg in J. van Heijenoort (ed.), From Frege to Gödel: A

Source Book in Mathematical Logic, 1879–1931, Cambridge, MA: Harvard University Press. 1967.

Frege, G. (1918-19), *Logical Investigations*. Yale University Press.

Frixione, M., Iaquinto, S., Vignolo, M. (2016), *Introduzione alle logiche modali*. Roma-Bari: Laterza.

Galvan, S. (1991), *Logiche intensionali. Sistemi proposizionali di logica modale, deontica, epistemica*. Milano: Franco Angeli.

Galvan, S., Mancosu, P., Zach R. (2021)*, An Introduction to Proof Theory. Normalization, Cut-Elimination, and Consistency Proofs*. Oxford: Oxford University Press.

Gettier, E. (1963), Is Justified True Belief Knowledge?, *Analysis*, 23, pp. 121–123.

Giere, R. N. (2004), How Models are Used to Represent Reality, *Philosophy of Science*, 71, pp. 742–752.

Gilhland-Swetland, A.J. (1998), Setting the stage. In Baca, M., ed. *Introduction to metadata: pathways to digital Information*. Los Angeles, CA: Getty Research Institute, pp. l–12.

Gruber Thomas, R. (1993), Toward Principles for the Design of Ontologies Used for Knowledge Sharing, *In International Journal Human-Computer Studies*, 43, pp. 907–928.

Gruber, T. R. (2009), Ontology. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems*. Springer.

Guarino, N., Musen, M. (2015), Applied ontology: The next decade begins, *Applied Ontology*, 10, pp. 1–4.

Hintikka, J. (1957), *Quantifiers in Deontic Logic*. Societas Scientiarum Fennica, Commentationes humanarum litterarum, 23, Helsingfors.

Hintikka, J. (1961), Modality and Quantification, *Theoria*, 27, pp. 119–28.

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.

Hjørland, B. (2008), What is Knowledge Organization (KO)?, *Knowledge Organization*, 35, No.2/No.3.

Hjørland, B. (2011), The Importance of theories of knowledge: Indexing and information retrieval as an example. *Journal of the American Society for Information Science and Technology*, 62:1, pp. 72–77.

Hjørland, B. (2013), Facet analysis: The logical approach to knowledge organization. *Information Processing & Management*, 49, pp. 545–557.

Hjørland, B. (2016), *Knowledge organization***,** in Encyclopedia of Knowledge Organization, eds. B. Hjørland – C. Gnoli.

Hu, Y., Hang Li, Yunbo Cao, Li Teng, Meyerzon, D., Zheng, Q. (2005), Automatic extraction of titles from general documents using machine learning, *Information Processing & Management*, 42:5, pp. 1276–1293. https://doi.org/10.1016/j.ipn2005.12.001.

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.*

Jashapara, A. (2004), *Knowledge management: an integrated approach*. Harlow: Financial Times Prentice Hall.

Kanger, S. (1957a), *Provability in Logic*. Dissertation, University of Stockholm.

Kanger, S. (1957b), A Note on Quantification and Modalities, *Theoria*, 23, pp. 131–134.

Kedrov, B.M. (1975), *Klassifizierung der Wissenschaften. Zwei Bände*, Köln: Pahl-Rugenstein.

Kornyshova, E., Deneckère, R., and Salinesi, C. (2007), *Method Chunks Selection by Multicriteria Techniques: an Extension of the Assembly-based Approach, Situational Method Engineering (ME)*. Geneva, Switzerland.

Kripke, S. (1959), A Completeness Theorem in Modal Logic, *Journal of Symbolic Logic*, 24, pp. 1–14.

Kripke, S. (1963), Semantical Considerations on Modal Logic, *Acta Philosophica Fennica*, 16, pp. 83–94.

Lowe, E.J. (2006), *The Four-Category Ontology. A Metaphysical Foundation for Natural Science*. Oxford: Oxford University Press.

Lozano, M.G., Brynielsson, J., Franke, U., Rosell, M., Tjörnhammara, E., Varga, S., Vlassov, V. (2020), Veracity assessment of online data, *Decision Support Systems*.

Lukoianova T., Rubin, V.L., (2014), Veracity Roadmap: Is Big Data Objective, Truthful and Credible?, *Advances in Classification Research Online*, pp. 4–15.

Martin, B. (2008), Knowledge management. In: Cronin, B., ed., *Annual Review of Information Science and Technology,* Volume 42, pp 371–424.

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

Minker, J. (2000), *Logic-Based Artificial Intelligence*. Springer.

NISO, (2004), *Understanding metadata*, 4733 Bethesda Avenue, Suite 300, Bethesda, MD 20814 USA: NISO.

Ngo-The, A., Ruhe, G. (2005), Decision Support in Requirements Engineering, *Engineering and Managing Software Requirements*, Ed. By A. Aurum and C. Wohlin, pp. 267–286.

Plato, (2017), *Euthyphro, Apology, Crito, Phaedo*. Greek with translation by Chris Emlyn-Jones and William Preddy. Loeb Classical Library 36. Harvard University Press.

Plato, (1921), *Theaetetus, Sophist*. Greek with translation by H. N. Fowler. Loeb Classical Library 123. Harvard University Press.

Pomerantz, J. (2015), *Metadata*. Cambridge, MA: MIT Press.

Post, E.L. (1921), Introduction to a General Theory of Elementary Propositions, *American Journal of Mathematics*, 43, pp. 163–185.

Priest, G. (2008), *An introduction to non-classical logic: From if to is* (2nd ed.), Cambridge: Cambridge University Press.

Rollett, H., (2003), *Knowledge management: processes and technologies*. Boston, MA: K1uwer Academic Publishers.

Roy, B. (2005), Paradigms and challenges, Book chapter, *In Multiple Criteria Decision Analysis - State of the Art Survey*, Springer. editor(s) J. Figueira, S. Greco, M. Ehrgott, pp. 3–24.

Ruhe, G. (2003), Software Engineering Decision Support – Methodology and Applications. In: Innovations in Decision Support Systems (Ed. by Tonfoni and Jain), *International Series on Advanced Intelligence*, Volume 3, pp. 143–174.

Runggaldier, E., Kanzian, C. (1998). *Grundprobleme der analytischen ontologie*. Verlag Ferdinand Schöning.

Samurin, E.I. (1964), *Geschichte der bibliothekarisch-bibliographischen Klassifikation*, Verlag Dokumentation.

Santos, Y.D. (2020), A Four-Valued Dynamic Epistemic Logic, *Journal of Logic, Language and Information*, 29, pp. 451–489.

Schaffer, J. (2009), On what grounds what. In David Manley, David J. Chalmers & Ryan Wasserman (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*. Oxford University Press, pp. 347–383.

Simon, H. (1960), *The New Science of Management Decision*. Harper&Row.

Smith, N. J. J. (2012), *Logic: the laws of truth*. Princeton University Press.

Smullyan, R. M. (1968), *First-order logic*. Berlin: Springer-Verlag.

Snow, D. (2012). Adding a 4th V to BIG Data – Veracity, http://dsnowondb2.blogspot.se/ 2012/07/adding-4th-v-tobig-data-veracity.html, 2012.

Tahko, T.E. (2023), The modal basis of scientific modelling, *Synthese*, 201:75, pp. 1–16.

Tambassi, T. (2022), On the Informativeness of Information System Ontologies. *Philosophia.*

Taylor, A.G, (2004), *The organization of information*. Westport, Conn: Libraries Unlimited.

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*.

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, 2015.

Turbanti, G. (2020), *Logica e mondi possibili*. Pisa: Pisa University Press.

van Ditmarsch, H., Halpern, J., van Der Hoek, W., Kooi, B. (2015), *Handbook of Epistemic Logic*. College Publications.

van Inwagen, P. (1998), *Meta-Ontology*, Erkenntnis, 48, pp. 223–250.

Varzi, A. C. (2011), On the boundary between material and formal ontology. In B. Smith, R. Mizoguchi, & S. Nakagawa (Eds.), *Interdisciplinary ontology* (Vol. 3, pp. 3–8), Keio University.

Weisberg, M. (2013), *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Williamson, T. (2013), Gettier Cases in Epistemic Logic, *Inquiry: An Interdisciplinary Journal of Philosophy*, 56, pp. 1– 14.

Wittgenstein, L. (1921), *Tractatus Logico-Philosophicus*. C. K. Ogden (trans.), London: Routledge & Kegan Paul. 1922. Originally published as “LogischPhilosophische Abhandlung”, in Annalen der Naturphilosophische, XIV (3/4), 1921.