FOL-BASED APPLIED ONTOLOGY FOR METADATA EXTRACTION IN MATHEMATICAL KNOWLEDGE MANAGEMENT

SIMONE CUCONATO

Abstract. Mathematical Knowledge Management (MKM) is a recent interdisciplinary field of research in the intersection of mathematics, computer science, data science, knowledge engineering and library science. The main goal of this paper is to build a first-order logic (FOL)-based applied ontology for metadata extraction, by the name MADME (MAke Decision for Metadata Extraction). The MADME procedure with its mathematical logic ontology is able to determine the best metadata extraction systems from heterogeneous digital mathematical documents, and to support the research community in MKM.

Mathematics Subject Classification (2010): 03B10, 03B70, 68V30, 68V35.
Key words: Mathematical logic, formal ontology, data mining.

Article history:
Received: September 27, 2023
Received in revised form: November 27, 2023
Accepted: November 28, 2023

1. Introduction

Mathematical Knowledge Management (MKM) is a recent interdisciplinary field of research that brings together mathematicians, computer scientists, engineers, and digital humanists [3]. While MKM is indeed a new field of research, mathematicians have been engaged in the management of mathematical knowledge for hundreds, if not thousands, of years. The history of MKM goes back much further at least to Euclid's great and extraordinarily influential *Elements*. Multiple factors have contributed to a heightened interest in the management of mathematical knowledge: (i) since World War II there has been an explosion in the mathematical knowledge produced by mathematicians; (ii) simultaneously, there has been a corresponding surge in mathematical knowledge generated as a byproduct of the work conducted by scientists and engineers; (iii) with the increasing prevalence of computer and communication systems, there is currently a significant transformation underway in the management of mathematical knowledge, encompassing its articulation, organization, dissemination, and accessibility.

In this paper, a first-order logic (FOL)-based decision-making ontology for metadata extraction from heterogeneous mathematical document sources will be presented[1]. In data science and knowledge engineering the majority of scholars have made a clear distinction between data, information and knowledge. 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

[1]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 [1, 12]. Furthermore, in recent years there have been increasing applications of logic to data science [4, 5].

acted upon). Information or knowledge that is organized, stored, managed or shared requires a particular type of meta-information or meta-knowledge: metadata[2]. Metadata emphasize meta-information or meta-knowledge aspects in that they describe the content, quality, condition, and other characteristics of other data or information. Our logical-ontological procedure, by the name MADME (MAke Decision for Metadata Extraction), moves at the method and model level, and thanks to the metadata extracted is able to support mathematical knowledge management. Given a digital mathematical document, the main objective of the MADME procedure is to develop a mathematical decision-making (DM)[3] ontology that can guide the choice of metadata extraction systems. The MADME procedure includes three elements: DM ontology (DMO), DM rules (DMR), and DM procedure (DMP). DMO is an informal and formal representation of digital objects. DMR is derived from DMO and are formal rules written in FOL that define all decision steps in detail. The DMP provides a set of methodological guidelines for the application of DMR. This paper is organized as follows. First, in Section 2, we lay down some basic preliminaries, and we define the fundamental concepts of our ontology engineering. Then, in section 3 we provide the mathematical basis of MADME in FOL, and the sixty-eight decision rules underlying logical calculus for metadata extraction. In Section 4 we give examples of possible applications of MADME procedure, before concluding.

2. Informal decision-making ontology

In philosophical contexts [2],"ontology" has traditionally been defined as the theory of what exists (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[4]. In recent times, the use of the term "ontology" has become prominent in computer science [14], engineering [6] and information science [9, 17]. Tom Gruber [8] 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 et. al [15] considers an ontology as a way of representing a common understanding of a domain.

In this paper, our definition of "ontology" is the following:

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

Given the application of our ontology to heterogeneous mathematical document sources, and since our document sources will be digital sources, our ontology will have a particular category of objects: digital objects. In general, a digital object is defined as "an object composed of a set of bit sequences"[5]. In our domain a digital object is based on three fundamental classes or categories that cannot be reduced to anything else: digital mathematical documents, metadata extraction systems, and metadata sets.

Documents class (D) are divided into four subclasses: text (D_t), images (D_i), audio (D_a), video (D_v). The metadata extraction systems class (S) is divided into four subclasses: from text (S_{from_t}), from images (S_{from_i}), from audio (S_{from_a}), from video (S_{from_v}). The metadata sets class is divided into four subclasses: of text (M_{of_t}), of images (M_{of_i}), of audio (M_{of_a}), of video (M_{of_v}). Subclasses metadata extraction systems and metadata sets will also have specific instances of individuals. The subclasses of S will have the following metadata extraction systems as possible instances: Cermine (c), OCR++ (o),

[2]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.

[3]We define decision-making 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.

[4]It is important to point out that the meaning of "informal ontology" that we introduce in this section must be understood in the sense of ontology without logical-mathematical symbolism. In general, from a philosophical point of view, the ontology we are describing is already a formal ontology [10].

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

Grobid (g), Fits (f), Apache Tika (a), Emet (e), IPTC Photo Metadata (i), and Metadata Extractor (m). More in detail:

- S_{from_t}: c, o, g, f, a, e
- S_{from_i}: f, i, m, a, e
- S_{from_a}: f, a, e
- S_{from_v}: f, a, e, m

The subclasses of M will have the following metadata sets as possible instances: metadata sets of Cermine (mc), metadata sets of OCR++ (mo), metadata sets of Grobid (mg), metadata sets of Fits (mf), metadata sets of Apache Tika (ma), metadata sets of Emet (me), metadata sets of IPTC Photo Metadata (mi), and metadata sets of Metadata Extractor (mm). More in detail:

- M_{from_t}: mc, mo, mg, mf, ma, me
- M_{from_i}: mf, mi, mm, ma, me
- M_{from_a}: mf, ma, me
- M_{from_v}: mf, ma, me, mm

In addition, objects belonging to the document class will instantiate specific properties related to the format Φ. The format Φ 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.

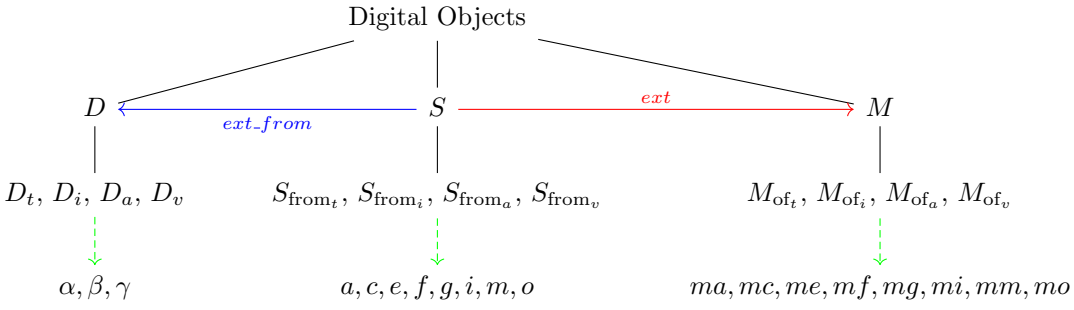
In particular, an object belonging to the subclass D_t can have a format of type PDF, DOC, DOCX, PAGES; to the subclass D_i, it can instantiate the formats BMP, GIF, JPEG, MP3, BFW, FLAC; to the subclass D_a, it can instantiate the format MP3, and to the subclass D_v, it can instantiate the formats BFW, MP4. Given an object x belonging to class D, we write Φx to denote the fact that x instantiates a property of the format Φ. For example:

- $\Phi_{\text{PDF}}x$ stands for "x is PDF"[6]

Additionally, we have two relations between classes:

- S ext M stands for "S extracts M"
- S ext_from D stands for "S extracts from D"

Therefore, the informal ontology described in this section serves us primarily as a design ontology:



Where:

- D standing for "the class of documents"
- S standing for "the class of metadata extraction systems"

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

- M standing for "the class of metadata sets"
- D_t, D_i, D_a, D_v are subclasses of the documents class D
- S_{from_t}, S_{from_i}, S_{from_a}, S_{from_v} are subclasses of the metadata extraction systems class S
- M_{of_t}, M_{of_i}, M_{of_a}, M_{of_v} are subclasses of the metadata sets class M
- α, β, γ[7] indicate the input instance of a mathematical digital document
- a, c, e, f, g, i, m, o are the possible metadata extraction systems instances
- $ma, mc, me, mf, mg, mi, mm, mo$ are the possible metadata sets instances
- the blue arrow represents the relationship: S ext M
- the red arrow represents the relationship: S ext_from D
- the green arrow indicates specific instances

3. The madme procedure

DMO can be defined as a heavyweight ontology, since it includes classes, subclasses, relationships between classes, instances, axioms, constraints and, especially, decision rules. This section will provide the mathematical basis in the language of first-order logic of the main notions and relations presented in the Section 2, and the sixty-eight decision rules underlying logical calculus[8] for metadata extraction. 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 exclusive disjunctions between classes:

$$D : \{x \mid x \text{ is a document}\}$$
$$(A1)\ \forall x(x \in D \leftrightarrow (x \notin S \wedge x \notin M) \wedge (x \in D_t \veebar x \in D_i \veebar x \in D_a \veebar x \in D_v))$$

$$S : \{x \mid x \text{ is a metadata extraction system}\}$$
$$(A2)\ \forall x(x \in S \leftrightarrow (x \notin D \wedge x \notin M) \wedge (x \in S_{from_t} \veebar x \in S_{from_i} \veebar x \in S_{from_a} \veebar x \in S_{from_v}))$$

$$M : \{x \mid x \text{ is a metadata set}\}$$
$$(A3)\ \forall x(x \in M \leftrightarrow (x \notin D \wedge x \notin S) \wedge (x \in M_{of_t} \veebar x \in M_{of_i} \veebar x \in M_{of_a} \veebar x \in M_{of_v}))$$

Axiom (A1) states that an object belongs to class D if and only if it does not belong to class S or M and belongs to subclass D_t or D_i or D_a or D_v. Axiom (A2) states that an object belongs to class S if and only if it does not belong to class D or M and belongs to subclass S_{from_t} or S_{from_i} or S_{from_a} or S_{from_v}. Lastly, axiom (A3) states that an object belongs to class M if and only if it does not belong to class D or S and belongs to subclass M_{of_t} or M_{of_i} or M_{of_a} or M_{of_v}.

The document class D will also have these specific axioms:

$$(A4)\ \forall x(x \in D \rightarrow \Phi x)$$
$$(A5)\ \forall x(x \in D_t \rightarrow \Phi_{PDF} x \veebar \Phi_{DOC} x \veebar \Phi_{DOCX} x \veebar \Phi_{PAGES} x)$$

[7] We will use the Greek letters α, β, γ for enunciative metavariables.

[8] The origins of the classical logic, as it was, and still often is called, go back to antiquity and are due to giants of Western thought such as Plato and Aristotle. 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 George Boole, who is sometimes regarded as the founder of mathematical logic. The classical calculus was first formulated as a formal axiomatic system by the eminent German logician Gottlob Frege, while the *sequent calculus*, first introduced by Gerhard Gentzen, is arguably the most elegant and flexible system for writing proofs. Today, modern logic is a formal, symbolic system that tries to discern *the laws of truth* [16]. In general, the job of describing a logical system comes in three parts: syntax, semantic and axiomatic. Syntax studies the composition of the propositions of a language, semantic studies the logical consequence relation from a semantic point of view, and axiomatic studies the logical consequence relation from a syntactic point of view. This paper is inspired by the logical rigour of *proof theorists* [11].

$$(A6)\ \forall x(x \in D_i \rightarrow \Phi_{BMP}x \veebar \Phi_{GIF}x \veebar \Phi_{JPEG}x \veebar \Phi_{MP3}x \veebar \Phi_{BFW}x \veebar \Phi_{FLAC}x)$$
$$(A7)\ \forall x(x \in D_a \rightarrow \Phi_{MP3}x)$$
$$(A8)\ \forall x(x \in D_v \rightarrow \Phi_{BFW}x \veebar \Phi_{MP4}x)$$

Axiom (A4) states that for all object x belonging to class D then x instantiates a property of the format Φ. Axioms (A5), (A6), (A7), and (A8), on the other hand, determine, based on the subclass to which an object x belongs, which formats Φ that object can instantiate.

The constraint (C1) below follows from (A2) and our DMO:

$$(C1)\ \forall x(x \in S \rightarrow (x = c \veebar x = o \veebar x = g \veebar x = f \veebar x = a \veebar x = e \veebar x = i \veebar x = m))$$

The constraint (C2) below follows from (A3) and our DMO:

$$(C2)\ \forall x(x \in M \rightarrow (x = mc \veebar x = mo \veebar x = mg \veebar x = mf \veebar x = ma \veebar x = me \veebar x = mi \veebar x = mm))$$

Some axioms on the relationships between classes are now introduced:

S ext M standing for "S extracts M":
$$(A9)\ S \text{ ext M} \rightarrow \exists x \exists y(x \in S \wedge y \in M)$$

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

S ext_from D standing for "S extracts from M":
$$(A10)\ S \text{ ext\_from D} \rightarrow \exists x \exists y(x \in S \wedge y \in D)$$
$$(A11)\ \forall x(x \in S \rightarrow (\exists y(y \in D \wedge x \text{ ext\_from } y) \veebar \neg \exists y(y \in D \wedge x \text{ ext\_from } y)))$$

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

The constraint (C3) below follows from (A9) and (A10):

$$(C3)\ x \text{ ext\_from } y \rightarrow \exists z(z \in M \wedge x \text{ ext } z)$$

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

(R1) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = c \wedge y \text{ ext\_from } x))$
(R2) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = o \wedge y \text{ ext\_from } x))$
(R3) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = g \wedge y \text{ ext\_from } x))$
(R4) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R5) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R6) $\forall x(x \in D_t \wedge \Phi_{PDF}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R7) $\forall x(x \in D_t \wedge \Phi_{DOC}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R8) $\forall x(x \in D_t \wedge \Phi_{DOCX}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R9) $\forall x(x \in D_t \wedge \Phi_{PAGES}x \rightarrow \neg \exists y(y \in S \wedge y \text{ ext\_from } x))$
(R10) $\exists x(x \in S \wedge x = c \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mc \wedge x \text{ ext } z)$
(R11) $\exists x(x \in S \wedge x = o \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mo \wedge x \text{ ext } z)$
(R12) $\exists x(x \in S \wedge x = g \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mg \wedge x \text{ ext } z)$
(R13) $\exists x(x \in S \wedge x = f \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mf \wedge x \text{ ext } z)$
(R14) $\exists x(x \in S \wedge x = a \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = ma \wedge x \text{ ext } z)$
(R15) $\exists x(x \in S \wedge x = e \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = me \wedge x \text{ ext } z)$
(R16) $\forall x(x \in D_i \wedge \Phi_{BMP}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R17) $\forall x(x \in D_i \wedge \Phi_{BMP}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R18) $\forall x(x \in D_i \wedge \Phi_{BMP}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R19) $\forall x(x \in D_i \wedge \Phi_{BMP}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R20) $\forall x(x \in D_i \wedge \Phi_{BMP}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R21) $\forall x(x \in D_i \wedge \Phi_{GIF}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$

(R22) $\forall x(x \in D_i \wedge \Phi_{GIF}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R23) $\forall x(x \in D_i \wedge \Phi_{GIF}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R24) $\forall x(x \in D_i \wedge \Phi_{GIF}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R25) $\forall x(x \in D_i \wedge \Phi_{GIF}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R26) $\forall x(x \in D_i \wedge \Phi_{JPEG}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R27) $\forall x(x \in D_i \wedge \Phi_{JPEG}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R28) $\forall x(x \in D_i \wedge \Phi_{JPEG}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R29) $\forall x(x \in D_i \wedge \Phi_{JPEG}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R30) $\forall x(x \in D_i \wedge \Phi_{JPEG}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R31) $\forall x(x \in D_i \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R32) $\forall x(x \in D_i \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R33) $\forall x(x \in D_i \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R34) $\forall x(x \in D_i \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R35) $\forall x(x \in D_i \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R36) $\forall x(x \in D_i \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R37) $\forall x(x \in D_i \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R38) $\forall x(x \in D_i \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R39) $\forall x(x \in D_i \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R40) $\forall x(x \in D_i \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R41) $\forall x(x \in D_i \wedge \Phi_{FLAC}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R42) $\forall x(x \in D_i \wedge \Phi_{FLAC}x \rightarrow \exists y(y \in S \wedge y = i \wedge y \text{ ext\_from } x))$
(R43) $\forall x(x \in D_i \wedge \Phi_{FLAC}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R44) $\forall x(x \in D_i \wedge \Phi_{FLAC}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R45) $\forall x(x \in D_i \wedge \Phi_{FLAC}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R46) $\exists x(x \in S \wedge x = f \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mf \wedge x \text{ ext } z)$
(R47) $\exists x(x \in S \wedge x = i \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mi \wedge x \text{ ext } z)$
(R48) $\exists x(x \in S \wedge x = m \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mm \wedge x \text{ ext } z)$
(R49) $\exists x(x \in S \wedge x = a \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = ma \wedge x \text{ ext } z)$
(R50) $\exists x(x \in S \wedge x = e \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = me \wedge x \text{ ext } z)$
(R51) $\forall x(x \in D_a \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R52) $\forall x(x \in D_a \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R53) $\forall x(x \in D_a \wedge \Phi_{MP3}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R54) $\exists x(x \in S \wedge x = f \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mf \wedge x \text{ ext } z)$
(R55) $\exists x(x \in S \wedge x = a \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = ma \wedge x \text{ ext } z)$
(R56) $\exists x(x \in S \wedge x = e \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = me \wedge x \text{ ext } z)$
(R57) $\forall x(x \in D_v \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R58) $\forall x(x \in D_v \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R59) $\forall x(x \in D_v \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R60) $\forall x(x \in D_v \wedge \Phi_{BFW}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R61) $\forall x(x \in D_v \wedge \Phi_{MP4}x \rightarrow \exists y(y \in S \wedge y = f \wedge y \text{ ext\_from } x))$
(R62) $\forall x(x \in D_v \wedge \Phi_{MP4}x \rightarrow \exists y(y \in S \wedge y = a \wedge y \text{ ext\_from } x))$
(R63) $\forall x(x \in D_v \wedge \Phi_{MP4}x \rightarrow \exists y(y \in S \wedge y = e \wedge y \text{ ext\_from } x))$
(R64) $\forall x(x \in D_v \wedge \Phi_{MP4}x \rightarrow \exists y(y \in S \wedge y = m \wedge y \text{ ext\_from } x))$
(R65) $\exists x(x \in S \wedge x = f \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mf \wedge x \text{ ext } z)$
(R66) $\exists x(x \in S \wedge x = a \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = ma \wedge x \text{ ext } z)$
(R67) $\exists x(x \in S \wedge x = e \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = me \wedge x \text{ ext } z)$
(R68) $\exists x(x \in S \wedge x = m \wedge x \text{ ext\_from } y) \rightarrow \exists z(z \in M \wedge z = mm \wedge x \text{ ext } z)$

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:

(1) Step 1. Evaluate the type of digital document source
   - Identify the class in D
   - Identify the format Φ
(2) Step 2. Apply decision rules.
(3) Step 3. Evaluate the extraction systems proposed by the procedure

4. Madme in action

This section will show how the MADME procedure makes its choices in practice. Specifically, the examples that will be considered will concern text documents, images, audio and video. However, only in the first example will we refer to a concrete text document.

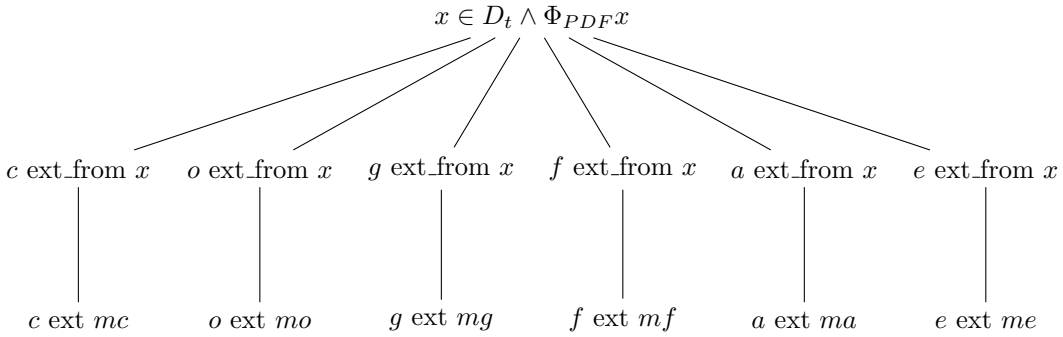
Example 4.1. Given a digital document source $x = [7]$, based on the MADME procedure, the first step is the evaluation of the type of digital document source. In this case:

i) $x \in D_t$
ii) $\Phi_{PDF}x$

For this reason, $x \in D_t \wedge \Phi_{PDF}x$ will be the premise. The second step is the application of decision rules:

1. $x \in D_t \wedge \Phi_{PDF}x$	Premise
2. $y_1 \in S \wedge y_1 = c$	1-R1
3. c ext_from x	1,2-R1
4. $y_2 \in S \wedge y_2 = o$	1-R2
5. o ext_from x	1,4-R2
6. $y_3 \in S \wedge y_3 = g$	1-R3
7. g ext_from x	1,6-R3
8. $y_4 \in S \wedge y_4 = f$	1-R4
9. f ext_from x	1,8-R4
10. $y_5 \in S \wedge y_5 = a$	1-R5
11. a ext_from x	1,10-R5
12. $y_6 \in S \wedge y_6 = e$	1-R6
13. e ext_from x	1,12-R6
14. $z_1 \in M \wedge z_1 = mc$	2,3-R10
15. c ext mc	14-R10
16. $z_2 \in M \wedge z_2 = mo$	4,5-R11
17. o ext mo	16-R11
18. $z_3 \in M \wedge z_3 = mg$	6,7-R12
19. g ext mg	18-R12
20. $z_4 \in M \wedge z_4 = mf$	8,9-R13
21. f ext mf	20-R13
22. $z_5 \in M \wedge z_5 = ma$	10,11-R14
23. a ext ma	22-R14
24. $z_6 \in M \wedge z_6 = me$	12,13-R15
25. e ext me	24-R15

Finally, the third step evaluates the metadata extraction systems proposed by the procedure. The MADME procedure allows establishing, based on axioms, constraints, and rules, that if the input document is $x = [7]$, the choice will fall on the following metadata extraction systems: CERMINE (step 3. and 15.), OCR++ (step 5. and 17.), GROBID (step 7. and 19.), FITS (step 9. and 21.), Apache Tika (step 11. and 23), and EMET (step 13. and 25.). The following graph summarises the results obtained from the application of the decision rules:



Now, when the procedure generates multiple choices, it is appropriate to operate according to the following principle[9]:
If given a document x there is a choice between different extraction systems, then choose, whenever possible, the one with the best metadata extraction percentages.
The principle allows us to optimize and maximize the chances of correctly extracting metadata by choosing the best possible extraction system. In this case, considering the results of the scientific literature [18], the principle will opt for CERMINE [19]. Therefore, given as input the document $x = [7]$, the output produced by CERMINE an XML file in the NLM JATS format:

```
▼<article xmlns:xlink="http://www.w3.org/1999/xlink">
 ▼<front>
    <journal-meta/>
  ▼<article-meta>
   ▼<title-group>
      <article-title>Twisted Neumann-Zagier matrices</article-title>
    </title-group>
   ▼<contrib-group>
    ▼<contrib contrib-type="author">
       <string-name>Stavros Garoufalidis</string-name>
       <email>stavros@mpim-bonn.mpg.de</email>
       <xref ref-type="aff" rid="aff0">0</xref>
       <xref ref-type="aff" rid="aff1">1</xref>
     </contrib>
    ▼<contrib contrib-type="author">
       <string-name>Seokbeom Yoon</string-name>
       <xref ref-type="aff" rid="aff0">0</xref>
     </contrib>
    ▼<aff id="aff0">
       <label>0</label>
       <institution>International Center for Mathematics, Department of Mathematics, Southern University of Science and Technology</institution>
       ,
       <addr-line>Shenzhen</addr-line>
       ,
       <country country="CN">China</country>
     </aff>
    ▼<aff id="aff1">
       <label>1</label>
       <institution>Max Planck Institute for Mathematics</institution>
       ,
       <addr-line>Bonn</addr-line>
       ,
       <country country="DE">Germany</country>
     </aff>
   </contrib-group>
  ▼<pub-date>
     <year>2023</year>
   </pub-date>
   <fpage>2</fpage>
   <lpage>24</lpage>
  ▼<history>
   ▼<date date-type="accepted">
       <day>29</day>
       <month>7</month>
       <year>2023</year>
     </date>
    ▼<date date-type="received">
       <day>22</day>
       <month>12</month>
       <year>2022</year>
     </date>
   </history>
```

FIGURE 1. Partial XML of metadata extracted by CERMINE (extraction link)

Example 4.2. Given a digital document source x such that $x \in D_t \wedge \Phi_{PAGES} x$:
1. $x \in D_t \wedge \Phi_{PAGES} x$ Premise

[9]The principle is inspired by the famous *Ockham's razor*.

2. $\emptyset$ 1-R9

In the second example, the decision procedure is immediately blocked by rule (R9). In this case, the MADME procedure states that the set of possible choices is nothing other than the empty set $\emptyset$.

Example 4.3. Given a digital document source x such that $x \in D_i \wedge \Phi_{JPEG}x$, based on the MADME procedure $x \in D_i \wedge \Phi_{JPEG}x$ will be the premise:

1. $x \in D_i \wedge \Phi_{JPEG}x$ Premise
2. $y_1 \in S \wedge y_1 = f$ 1-R26
3. f ext_from x 1,2-R26
4. $y_2 \in S \wedge y_2 = i$ 1-R27
5. i ext_from x 1,4-R27
6. $y_3 \in S \wedge y_3 = m$ 1-R28
7. m ext_from x 1,6-R28
8. $y_4 \in S \wedge y_4 = a$ 1-R29
9. a ext_from x 1,8-R29
10. $y_5 \in S \wedge y_5 = e$ 1-R30
11. e ext_from x 1,10-R30
12. $z_1 \in M \wedge z_1 = mf$ 2,3-R46
13. f ext mf 12-R46
14. $z_2 \in M \wedge z_2 = mi$ 4,5-R47
15. i ext mi 14-R47
16. $z_3 \in M \wedge z_3 = mm$ 6,7-R48
17. m ext mm 16-R48
18. $z_4 \in M \wedge z_4 = ma$ 8,9-R49
19. a ext ma 18-R13
20. $z_5 \in M \wedge z_5 = me$ 10,11-R14
21. e ext me 20-R50

The MADME procedure allows establishing that if the input document $x \in D_i \wedge \Phi_{JPEG}x$ the choice will fall on the following extraction systems: FITS, IPTC, Metadata Extractor, and Apache Tika.

Example 4.4. Given a digital document source x such that $x \in D_a \wedge \Phi_{MP3}x$, based on the MADME procedure $x \in D_a \wedge \Phi_{MP3}x$ will be the premise:

1. $x \in D_a \wedge \Phi_{MP3}x$ Premise
2. $y_1 \in S \wedge y_1 = f$ 1-R51
3. f ext_from x 1,2-R51
4. $y_2 \in S \wedge y_2 = a$ 1-R52
5. a ext_from x 1,4-R52
6. $y_3 \in S \wedge y_3 = e$ 1-R53
7. e ext_from x 1,6-R53
8. $z_1 \in M \wedge z_1 = mf$ 2,3-R54
9. f ext mf 8-R54
10. $z_2 \in M \wedge z_2 = ma$ 4,5-R55
11. a ext ma 10-R55
12. $z_3 \in M \wedge z_3 = me$ 6,7-R56
13. e ext me 12-R56

The MADME procedure allows establishing that if the input document $x \in D_a \wedge \Phi_{MP3}x$ the choice will fall on the following extraction systems: FITS, Apache Tika, and EMET.

Example 4.5. Given a digital document source x such that $x \in D_v \wedge \Phi_{MP4}x$, based on the MADME procedure $x \in D_v \wedge \Phi_{MP4}x$ will be the premise:

1. $x \in D_v \wedge \Phi_{MP4}x$ Premise

2. $y_1 \in S \wedge y_1 = f$	1-R61
3. f ext_from x	1,2-R61
4. $y_2 \in S \wedge y_2 = a$	1-R62
5. a ext_from x	1,4-R62
6. $y_3 \in S \wedge y_3 = e$	1-R63
7. e ext_from x	1,6-R63
8. $y_4 \in S \wedge y_4 = m$	1-R64
9. m ext_from x	1,8-R64
10. $z_1 \in M \wedge z_1 = mf$	2,3-R65
11. f ext mf	10-R65
12. $z_2 \in M \wedge z_2 = ma$	4,5-R66
13. a ext ma	12-R66
14. $z_3 \in M \wedge z_3 = me$	6,7-R67
15. e ext me	14-R67
16. $z_4 \in M \wedge z_4 = mm$	8,9-R68
17. m ext mm	16-R68

The MADME procedure allows establishing that if the input document $x \in D_v \wedge \Phi_{MP4}x$ the choice will fall on the following extraction systems: FITS, Apache Tika, EMET, and Metadata Extractor.

5. Conclusion and future development

Ontologies have been used for the last decades for a set of tasks. Some of these tasks deal with interoperability issues and can be applied in different domains. In this paper we considered the domain of digital mathematical objects and used mathematical logic to develop formal proofs for a decision ontology for metadata extraction. A formal proof is not a natural language argument. It is a calculation that follows precise rules. The rules are *grounded* on formal notation and logical proof. Formal notation and proof are rigorous, unambiguous and can be checked mechanically. The strength of the MADME procedure is to determine the best metadata extracion systems from heterogeneous digital mathematical documents and, in this way, to support the research community in mathematical knowledge management. A future line of research could be to implement the MADME procedure. The presence of a rigorous formal apparatus can facilitate the translation of procedural rules into a programming language and, consequently, the fully automated development of MADME.

References

[1] M. Ben-Ari, *Mathematical Logic for Computer Science,* Springer, 2012.

[2] F. Berto and M. Plebani, *Ontology and meta-ontology: A contemporary guide,* Bloomsbury, 2015.

[3] J. Carette and W.M. Farmer, *A rreview of mathematical knowledge management,* in: J. Carette, L. Dixon, C.S. Coen, S.M. Watt (Eds.), *Intelligent Computer Mathematics: Proceeding of CICM 2009*, Lecture Notes in Computer Science, Springer, Berlin-Heidelberg, 2009, 233-246.

[4] S. Cuconato, *Epistemic logic for metadata modelling from scientific papers on covid-19,* Science & Philosophy **9(2)** (2021), 83-96.

[5] S. Cuconato, *A four-valued epistemic logic for metadata modelling from medical articles on pain therapies,* in: A.K. Das, J. Nayak, B. Naik, S. Vimal, D. Pelusi (Eds.), *Computational Intelligence in Pattern Recognition: Proceeding of CIPR 2023*, Lecture Notes in Networks and Systems, Springer, Singapore, 2023, 631-640.

[6] C. Roussey, F. Pinet, M.A. Kang and O. Corcho, *An introduction to ontologies and ontology engineering,* in G. Falquet et al. (Eds.), *Ontologies in Urban Development Projects. Advanced Information and Knowledge Processing*, Springer, London, 2011.

[7] S. Garoufalidis and S. Yoon, *Twisted Neumann–Zagier matrices,* Research in the Mathematical Sciences **10(37)** (2023), 1-23.

[8] T.R. Gruber, *Toward principles for the design of ontologies used for knowledge sharing,* International Journal Human-Computer Studies **43(5-6)** (1995), 907-928.
[9] N. Guarino and M. Musen, *Applied ontology: The next decade begins,* Applied Ontology **10(1)** (2015), 1-4.
[10] J. Hakkarainen and M. Keinänen, *Formal Ontology,* Cambridge University Press, 2023.
[11] P. Mancosu, S. Galvan and R. Zach, *An Introduction to Proof Theory: Normalization, Cut-Elimination, and Consistency Proofs*, Oxford University Press, 2021.
[12] J. Minker, *Logic-Based Artificial Intelligence,* Springer, 2012.
[13] H. Rollett, *Knowledge Management: Processes and Technologies,* Springer, 2003.
[14] A.A Salatino, T. Thanapalasingam, A. Mannocci, A. Birukou, F. Osborne and E. Motta, *The computer science ontology: A comprehensive automatically-generated taxonomy of research areas,* Data Intelligence **2(3)** (2020), 379-416.
[15] D.M. Sánchez, J.M. Cavero and E.M. Martínez, *The road toward ontologies,* in: R. Sharman, R. Kishore, R. Ramesh (Eds.), *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*, Springer, New York, 2007.
[16] N.J.J. Smith, *Logic: The Laws of Truth,* Princeton University Press, 2012.
[17] T. Tambassi, *On the informativeness of information system ontologies,* Philosophia **50(5)** (2022), 2675-2684.
[18] D. Tkaczyk, P. Szostek, P. Jan Dendek, M. Fedoryszak and L. Bolikowski, *Cermine–automatic extraction of metadata and references from scientific literature,* 11th IAPR International Workshop on Document Analysis Systems (2014), 217–221.
[19] D. Tkaczyk, P. Szostek, P. Jan Dendek, M. Fedoryszak and L. Bolikowski, *Cermine–automatic extraction of metadata and references from scientific literature,* International Journal on Document Analysis and Recognition (IJDAR), Springer **18(4)** (2015), 317-335.

Department of Computer Engineering, Modeling, Electronics and Systems Engineering, University of Calabria, Cosenza, Italy
Email address: simone.cuconato@unical.it