Data Mining Meets Logic: Situation-Based Modal Logic and Metadata Veracity

Simone Cuconato*

Abstract

Logic and theoretical computer science are deeply interconnected, with logic forming a foundational pillar in the emergence of computer science. This connection has grown stronger over time, driven by advancements in symbolic systems within artificial intelligence, formal verification methods, and automated reasoning techniques. However, within the vast landscape of computer science, data science remains an area where the link to logic is relatively underdeveloped. In this paper, I apply algebraic logic to address an urgent challenge for data scientists: *metadata veracity*. Inspired by Willard Van Orman Quine's well-known slogan, "*No entity without identity*", I propose, "*No data without metadata*", underscoring the importance of descriptive metadata in scientific articles. The central idea of my framework is that the geometric zones of a scientific article – with its atomic metadata – correspond to an algebraic situation space. Consequently, I develop a situation-based semantics grounded on the idea that situations are portions of a *possible world*, a *possible world* is a PDF document and that sentences point to situations and describe what is going on in them. Applying mathematical logic tools to data mining enables a rigorous framework for defining metadata veracity, offering a structured approach to assess the accuracy and reliability of extracted information.

Keywords: Modal logic; Situation semantics; Metadata veracity[1]

*University of Calabria, Via P. Bucci, Rende(CS) 87036, Italy & LUISS University, Rome, Italy; simone.cuconato@unical.it.

[1]Received on November 09, 2024. Accepted on May 04, 2025. Published on June 30, 2025. doi: 10.23756/sp.v13i1.1658. ISSN 2282-7757; eISSN 2282-7765.

1 Preliminaries

In recent years, one of the concepts that has captured the interest of data scientists is that of data quality, a complex and multidimensional concept, to the definition of which different dimensions concur: "Data quality is a multidimensional concept combining consistency, freshness, completeness, and accuracy of the data as a non-exhaustive list of dimensions to characterize data fitness-of-use" [1]. In the era of Big Data, data quality has become more important than ever. Normally, volume, velocity and variety are used to characterise the main features of Big Data [2]. However, the importance of the fourth 'V' of Big Data, *veracity*, is increasingly recognised. For the purposes of my paper, the conceptual dimensions I consider are those of data *completeness* and *consistency*. In general, the following definitions of completeness and consistency are used in data science:

- "Completeness, pertinence and relevance refer to the capability of representing all and only the relevant aspects of the reality of interest" [3].

- "Consistency, cohesion and coherence refer to the capability of data to comply without contradictions to all properties of the reality of interest, as specified in terms of integrity constraints, data edits, business rules and other formalisms" [3].

In my paper, I apply the concepts of data completeness and consistency to a specific data type: metadata[2]. Metadata are crucial in data science, as they establish "*a criterion of identity* for data" [4]. Inspired by Willard Van Orman Quine's well-known slogan, "*No entity without identity*" [5], I propose, "*No data without metadata*," underscoring the importance of descriptive metadata in scientific articles. Specifically, I focus on descriptive metadata extracted from scientific articles and use one of the most effective extraction systems available: Cermine [6], a framework designed and trained to extract information from PubMed. Cermine is an open source tool specializes in extracting metadata and content from scientific articles in PDF format. Its modular architecture utilizes both supervised and unsupervised machine learning techniques. In details, Cermine's extraction workflow takes a PDF file on the input and produces its geometric hierarchical representation in Trueviz format:

> In this representation, an article is a list of pages, each page contains a list of zones, each zone contains a list of lines, each line

[2] The idea of metadata can be traced back to antiquity, with its initial presence in early libraries; however, the first published use of the term 'metadata' in the sense of 'data about data' likely appears in the first edition of NASA's Directory Interchange Format Manual, published in 1988.

> contains a list of words, and finally each word contains a list of characters. Each structure element can be described by its text content and bounding box (a rectangle enclosing the element).The structure contains also the natural reading order for the elements on each level. Additionally, labels describing the role in the document are assigned to zones [6].

Subsequently, every zone is labelled with one of four general categories: metadata, references, body or other. Metadata zone classification classifies all metadata zones into specific metadata classes, and from metadata zone classification metadata extraction path extracts a rich set of atomic metadata. The output format of the extraction workflow is NLM JATS. JATS (Journal Article Tag Suite) describes a rich set of XML elements and attributes for the description of scientific publications. Documents in JATS format can store a large range of structured and descriptive metadata of the document (title, authors, affiliations, journal name, publication date, abstract, keywords, etc.). I focus on four specific atomic metadata: title, authors, journal name, and keywords. The core of my situational framework is that the zones of a scientific article – with its atomic metadata –, correspond to an algebraic situation space. Consequently, I develop a situation-based semantics grounded on the idea that situations are portions of a possible world, a possible world is a PDF document and that sentences point to situations and describe what is going on in them. Applying mathematical logic tools to data science enables a rigorous framework for defining metadata veracity, offering a structured approach to assess the accuracy and reliability of extracted information[3].

2 Metadata as Situations

My situation semantics is based on the notion of situation space developed by Alessandro Giordani [9]. Giordani characterizes: *i*) the notion of situation space in a way similar to the notion of state space used by John Langshaw Austin [10] for developing *truthmaker semantics*; and *ii*) the connection between situations and sentences, drawing inspiration from the work of Kit Fine [11, 12]. Additionally, in line with Giordani, I assume that there is no substantial difference between the *toy world* described below and a generic *possible world*. Consequently, I develop my situation-based semantics on the idea that situations are portions of a world (and more specifically of a 'document world') and that sentences point to situations

[3]For an application of modal logic to metadata extraction systems based on machine learning see [7], while for a study on logical language capable of investigating interactive voting decisions in multi-agent systems in liquid democracy see [8].

and describe what is going on in them. Algebraically, a situation space is a *join-semilattice* [13, 14]. A join-semilattice is a set of situations *Sit* equipped with an idempotent, commutative, associative binary operation $\sqcup$. *Sit* is partially ordered by the relation $s \leq s'$, which by definition means $s \sqcup s' = s'$. In this order, $s \leq s'$ is the least upper bound of s and s'. The largest element of *Sit*, if it contains one, is denoted by 1.

Definition 1. A *situation space* is a join-semilattice with top, $(Sit, 1, \sqcup)$.

Sit is a set of situations, 1 is the top element, so that $s \sqcup 1 = 1$, and $\sqcup$ returns the join of two situations. By definition the binary operation is

1. *idempotent*: $s \sqcup s = s$, for every $s \in Sit$
2. *commutative*: $s \sqcup s' = s' \sqcup s$, for every $s, s' \in Sit$
3. *associative*: $s \sqcup (s' \sqcup s'') = (s \sqcup s') \sqcup s''$, for every $s, s', s'' \in Sit$.

Let us consider the following toy world.

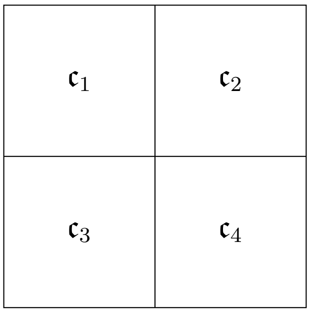

Figure 1: An algebraic toy world with four elementary cells.

This world consists of four elementary cells: c_1, c_2, c_3, c_4, numbered from the top-left to the bottom-right corner. A situation is any portion of the world obtained by joining certain cells together. For example, c_1, $c_1 \sqcup c_2$, and the entire world $1 = c_1 \sqcup c_2 \sqcup c_3 \sqcup c_4$ are situations. A moment's reflection reveals that, given there are 4 cells, the number of situations in this world is $2^4 - 1$[4]. The most relevant intuition of the situation semantics developed by Giordani concerns the way a sentence is related to the world[5]. Let p_i, $i = 1, \dots, 4$, be the sentence expressing the proposition that cell i of our toy world is filled, Giordani writes:

[4]Obviously, partial ordering on Sit is just spatial containment between the portions of the world constituted by the cells.

[5]Furthermore, this intuition is at the basis of John Austin's definition of truth.

> So, p_1 states that the first cell of the world is filled, and therefore it expresses a true proposition, while p_2 states that the fourth cell of the world is filled, thus expressing a false proposition. It makes perfect sense to assume that a sentence of this kind points to a situation and describes what is going on in it. Indeed, p_1 points to the first cell and describes it as filled, and similarly, p_2 points to the second cell and describes it as filled [9].

Graphically, I indicate in Fig. 2. a filled cell with the colour grey, and an unfilled cell with the colour white.

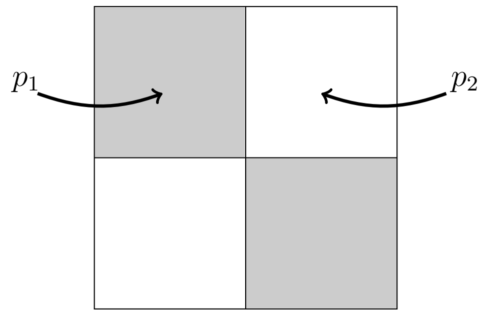

Figure 2: A toy world with filled and unfilled cells.

As for complex sentences, it makes sense to assume that $p_1 \wedge p_2$ points to the first two cells and describes them as a couple of filled cells, and therefore it is false. Similarly, $p_1 \vee p_2$ points again to the first two cells, but describes them as a couple of cells of which at least one is filled, thus being true. So, it makes sense to introduce a general assumption to the effect that complex sentences point to the cells described by their components. Finally, since $\top$ is a description of everything, it points to the whole world, and therefore any formula containing $\top$ points to the whole world as well.

In my application domain, the elementary cells of a *document world* correspond to the text zones of the document that contain the information related to the atomic metadata of interest. For example, in Fig. 3. the PDF document [15] from which I perform the metadata extraction in the example in Section 4., the title metadata zone is highlighted in red, the author metadata zone in blue, the journal metadata zone in yellow, and finally the keywords metadata zone in green. Metadata title zone with its atomic metadata correspond to cell $\mathfrak{c}_1$, metadata author to cell $\mathfrak{c}_2$, metadata journal to cell $\mathfrak{c}_3$, and finally, metadata keywords to cell $\mathfrak{c}_4$[6].

[6]My application domain primarily involves the extraction of structured, descriptive metadata from scientific articles in PDF format. In this setting, atomic metadata – such as the article title, author(s), journal, and keywords – are identified in distinct text zones of the document, each of

The Protein Journal (2023) 42:477–489
https://doi.org/10.1007/s10930-023-10137-1

Check for updates

On the Path to Optimal Alchemistry

Magnus Lundborg[1] · Jack Lidmar[2] · Berk Hess[3]

Accepted: 4 July 2023 / Published online: 31 August 2023
© The Author(s) 2023

Abstract
Alchemical free energy calculations have become a standard and widely used tool, in particular for calculating and comparing binding affinities of drugs. Although methods to compute such free energies have improved significantly over the last decades, the choice of path between the end states of interest is usually still the same as two decades ago. We will show that there is a fundamentally arbitrary, implicit choice of parametrization of this path. To address this, the notion of the length of a path or a metric is required. A metric recently introduced in the context of the accelerated weight histogram method also proves to be very useful here. We demonstrate that this metric can not only improve the efficiency of sampling along a given path, but that it can also be used to improve the actual choice of path. For a set of relevant use cases, the combination of these improvements can increase the efficiency of alchemical free energy calculations by up to a factor 16.

Keywords Alchemical free energy calculations · Lambda path optimization · Accelerated weight histogram method

1 Introduction

In alchemical free energy calculations the nature of the simulated particles is changed. The goal of classical alchemistry was to obtain more precious materials from cheaper ones. In alchemical free energy calculations the goal is to obtain the difference in free energy between systems of different composition or with different interactions. A single such free energy difference between different chemical states is not useful by itself. But when considering thermodynamic cycles, a closed loop of multiple of such free energy differences between multiple states, meaningful free energy differences can be computed, see e.g. [1]. The power of alchemical free energy calculations lies in the property that two (or more) well chosen alchemical transformations can often lead to much more efficient calculations than sampling the physical pathway between the states one is interested in. A typical example is the computation of the difference in binding free energy of two (or more) ligands to a protein by computing the difference in the free energy of mutating the ligand when bound to the protein and when in solution. This avoids sampling the, physical or unphysical, pathway of (un)binding the ligand and the protein. [2, 3]

Initially, the main use of alchemical free energy calculations was to parametrize and validate molecular mechanics force fields, in particular for proteins. [4–9] A common check is to compare the free energies of compounds in different solvents with experiment. Here the compound is (de)coupled from the solvent by scaling the interactions of the compound with the solvent. Over the last two decades drug binding free energies have become one of the main application areas of molecular simulations. [10–13] However, there are still few, if any, marketed drugs developed using molecular dynamics based binding free energy calculations. [14] Here one can distinguish absolute binding free energy calculations, which are in method similar to solvation free energy calculations but with additional restraints, and the relative free energy difference between binding of two different drugs to the same molecule or one drug to different molecules. Relative free energy differences are obtained using alchemical transformations of the ligand or binding partners. When the goal is to compute free energy differences of many rather similar molecules, this enables efficient free energy calculations by alchemical transformation of a few atoms for each edge in a graph of molecules.

✉ Berk Hess
hess@kth.se

[1] ERCO Pharma AB, 11439 Stockholm, Sweden

[2] Department of Physics, KTH Royal Institute of Technology, 10691 Stockholm, Sweden

[3] Department of Applied Physics, KTH Royal Institute of Technology, 10691, Stockholm, Science for Life Laboratory, Solna, Sweden

Springer

Figure 3: Document with highlighted metadata.

In addition, I denote with m the atomic extracted metadata, to wich I add a specific subscript according to whether it is a metadata title m_t, metadata author m_a, metadata journal m_j, and metadata keywords m_k. In this way, my language has only sentences expressed by a particular propositional formula $\ddot{p}$ having the form

$$\ddot{p} \equiv \nabla m \tag{1}$$

where ∇ indicates a particular predicate or properties that a metadata can instantiate (or have, or exemplify). Metaphysically speaking, a metadata must be

which is critical for establishing the document's identity and relevance in data indexing, retrieval, and analysis. These text zones align with specific cells in my framework, allowing logical and algebraic tools to assess metadata consistency and completeness within each zone.

understood as an object. I use the term 'object' "as applying to whatever bears properties" [16]. In more detail, a metadata may satisfy a particular predicate ∇ denoting the property in question or, equivalently, make true the corresponding sentences $\ddot{p}$. Metadata are objects, for they are *property-bearers*[7]. Formally, ∇ indicates the element's content of a specific metadata in the XML markup language produced by Cermine. For example, given a scientific article, if Cermine extracts the following XML as metadata author value

```
<contrib contrib-type="author">
  <string-name>Aristotle</string-name>
</contrib>
```

then $\nabla = Aristotle$ and therefore

$$\ddot{p} \equiv \textit{Aristotle}\ m_a \tag{2}$$

The propositional formula $\ddot{p}$ should read 'metadata author is $Aristotle$'. These aspects allow the development of intuition of how a sentence is related to the document world. Let $\ddot{p}_i$, $i = 1, \ldots, 4$, be the sentence expressing the proposition that cell i of my toy world instantiates (or has, or exemplifies) the atomic metadata information of the text zones of the PDF document of interest. Let us assume that, $\ddot{p}_2$ states that the second cell of the document world instantiates *Aristotle*, and therefore it expresses a false proposition and metadata is not correctly extracted (for this reason I coloured the cell red), while $\ddot{p}_3$ states that the third cell of the document world instantiates *The Protein Journal*, thus expressing a true proposition and a correct metadata extraction (for this reason I coloured the cell green). Also in my application domain "it makes perfect sense to assume that a sentence of this kind points to a situation and describes what is going on in it"[9]. Indeed, in Fig. 4. $\ddot{p}_2$ points to the second cell and describes it as containing the word '*Aristotle*' (not correctly) and, similarly, $\ddot{p}_3$ points to the third cell and describes it as containing the word '*The Protein Journal*' (correctly).

The final step is to define when an extraction is complete and consistent. A key challenge of data quality arises from the fact that the extracted information may exhibit bias, noise, outdatedness, inaccuracies, and misleading elements, rendering it unreliable. For this reason, I adapt the general definitions of completeness and consistency to the domain of metadata extracted from scientific articles:

- *Completeness* refers to the capability of a specific metadata extraction to represent all and only true propositional formulas $\ddot{p}$ about the *relevant aspects* of the document;

- *Consistency* refers to the capability of atomic metadata to comply without contradictions to all *document properties*.

[7] For a Meinongian interpretation of mathematical objects see [17**?**].

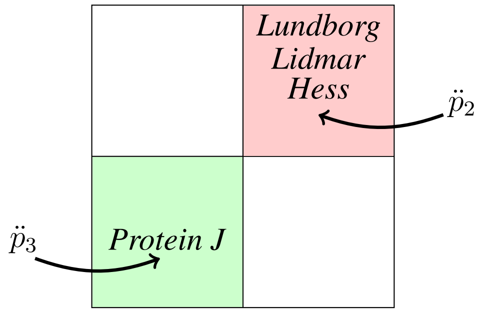

Figure 4: $\ddot{p}_2$ points to the second cell, and $\ddot{p}_3$ points to the third cell.

Where by 'relevant aspects' I mean that the atomic metadata bears exactly the relevant part of the text of the document without errors, e.g. the title not being given in full or with additions, while by 'document properties' I mean that the atomic metadata does not bear information falling under another class of metadata, e.g. that in the abstract metadata no information falling under the keywords metadata is also added.

3 Syntax and Semantic

Given these general theoretical issues, my basic language $\ddot{\mathcal{L}}$ consists of a non-empty set $\ddot{\Omega}$ of propositional variables, with a characteristic member of $\ddot{\Omega}$ being $\ddot{p}_i$ (where i is some natural number)[8] and the following logical constants: $\top$ (top); $\neg$ (negation); $\wedge$ (conjunction); $\vee$ (disjunction); $\rightarrow$ (classical implication); and the brackets[9]. The set $Fm(\ddot{\mathcal{L}})$ of formulas of $\ddot{\mathcal{L}}$ is defined inductively according to the following grammar:

$$\ddot{\phi}, \ddot{\psi} := \ddot{p} \mid \top \mid \neg\ddot{\phi} \mid (\ddot{\phi} \wedge \ddot{\psi}) \mid (\ddot{\phi} \vee \ddot{\psi}) \mid (\ddot{\phi} \rightarrow \ddot{\psi}) \tag{3}$$

where $\ddot{p} \in \ddot{\Omega}$.

As I have already specified, my semantic framework is a development of possible worlds semantics: "the content of a sentence is defined both in terms to the set of worlds at which it is true and in terms of the situation it points to" [9]. For this reason, the notion of a model for $\ddot{\mathcal{L}}$ is defined based on the notion of a situation frame.

Definition 2. A *situation frame* for $\ddot{\mathcal{L}}$ is a tuple (W, R, S) where

[8]Given our application domain, $\ddot{\Omega}$ is finite (therefore, for some n, $\ddot{\Omega} = \{\ddot{p}_1, \ddot{p}_2, \ldots, \ddot{p}_n\}$). I use the variable $\ddot{p}$ to range over the elements of $\ddot{\Omega}$.

[9]I remove the outermost parentheses and assume that conjunction and disjunction bind stronger than the conditional.

- $W \neq \emptyset$;
- $R : W \to \mathcal{P}(W)$ is such that $w \in R(w)$;
- $S : W \to$ **S** is such that S_w is a situation space $(\mathrm{Sit}_w, 1_w, \sqcup_w)$.

Intuitively, W is a set of possible worlds, while R is a function that associates each world $w \in W$ with the set $R(w)$ of worlds that are accessible from it. It is important specify that (W, R) coincides with a modal frame with a *reflexive relation*. S is a function that associates each world $w \in W$ with a situation space. In general, "the underlying idea is that every propositional variable is viewed as a device that points to a certain situation at a certain world and describes the way things are in that situation" [9]. This idea can be made more precise through a specific reference function for sentences.

Definition 3. A reference function σ for a frame for $\ddot{\mathcal{L}}$ is a function $\sigma : W \times Fm(\ddot{\mathcal{L}}) \to \bigcup_{w \in W} Sit_w$ such that

$$\sigma_w(\ddot{p}) \in Sit_w; \tag{4}$$

$$\sigma_w(\ddot{\phi}) = 1_w, \text{ if } \top \text{ occours in } \ddot{\phi}; \tag{5}$$

$$\sigma_w(\ddot{\phi}) = \sqcup_w \{\sigma_w(\ddot{p}) : \ddot{p} \in \ddot{\Omega}(\ddot{\phi})\}, \text{ if } \top \text{ does not occour in } \ddot{\phi}. \tag{6}$$

In this way, "the referent of a propositional variable at a world is the situation it points to at that world, while the referent of a complex formula at a world is the join of the referents of its components at that world" [9].

Definition 4. A *situation model* for $\ddot{\mathcal{L}}$ is a tuple (W, R, S, σ, V) where

- (W, R, S) is a situation frame;
- σ is a reference function for (W, R, S);
- $V : W \times \ddot{\Omega} \to \{1, 0\}$ is a valuation function.

V associates each pair constituted by a world $w \in W$ and a propositional variable $\ddot{p} \in \ddot{\Omega}$ with a truth value, the truth value of $\ddot{p}$ at w.

Definition 5. Truth of a formula at world in a model for $\ddot{\mathcal{L}}$. Let M be a situation model for $\ddot{\mathcal{L}}$. The notion of truth is defined so that

$$M, w \models \top \tag{7}$$

$$M, w \models \ddot{p} \quad \text{iff } V_w(\ddot{p}) = 1 \tag{8}$$

$$M, w \models \neg\ddot{\phi} \quad \text{iff } M, w \not\models \ddot{\phi} \tag{9}$$

$$M, w \models \ddot{\phi} \wedge \ddot{\psi} \quad \text{iff } M, w \models \ddot{\phi} \text{ and } M, w \models \ddot{\psi} \tag{10}$$

$$M, w \models \ddot{\phi} \vee \ddot{\psi} \quad \text{iff } M, w \models \ddot{\phi} \text{ or } M, w \models \ddot{\psi} \tag{11}$$

$$M, w \models \ddot{\phi} \rightarrow \ddot{\psi} \quad \text{iff } M, w \not\models \ddot{\phi} \text{ or } M, w \models \ddot{\psi} \tag{12}$$

Definition 6. The relation $\Vdash_{\ddot{\mathcal{L}}}$ of *logical consequence* for $\ddot{\mathcal{L}}$ is defined so that
$\ddot{\Delta} \Vdash_{\ddot{\mathcal{L}}} \ddot{\phi}$ iff, for all M and w, $M, w \models \ddot{\Delta}$ only if $M, w \models \ddot{\phi}$
where $\ddot{\Delta} \subseteq \ddot{\mathcal{L}}$ and $M, w \models \ddot{\Delta}$ iff $M, w \models \ddot{\delta}$ for all $\ddot{\delta} \in \ddot{\Delta}$.

System $(\ddot{\mathcal{L}}, \Vdash_{\ddot{\mathcal{L}}})$ is my basic semantic system of situation-based modal logic.

4 An Application Case

In this section, I analyze how my logical system can be applied to data mining. Given the PDF [15] and focusing only on the metadata title ($\nabla_1 m_t \equiv \ddot{p}_1$), authors ($\nabla_2 m_a \equiv \ddot{p}_2$), journal name ($\nabla_3 m_j \equiv \ddot{p}_3$), and keywords ($\nabla_4 m_k \equiv \ddot{p}_4$), Cermine extracts the following metadata.


```
▾<article xmlns:xlink="http://www.w3.org/1999/xlink">
  ▾<front>
    ▾<journal-meta>
      ▾<journal-title-group>
          <journal-title>The Protein Journal</journal-title>
        </journal-title-group>
      </journal-meta>
    ▾<article-meta>
      ▾<title-group>
          <article-title>On the Path to Optimal Alchemistry</article-title>
        </title-group>
      ▾<contrib-group>
        ▾<contrib contrib-type="author">
            <string-name>Magnus Lundborg</string-name>
            <xref ref-type="aff" rid="aff0">0</xref>
            <xref ref-type="aff" rid="aff1">1</xref>
            <xref ref-type="aff" rid="aff2">2</xref>
          </contrib>
        ▾<contrib contrib-type="author">
            <string-name>Jack Lidmar</string-name>
            <xref ref-type="aff" rid="aff0">0</xref>
            <xref ref-type="aff" rid="aff1">1</xref>
            <xref ref-type="aff" rid="aff2">2</xref>
          </contrib>
        ▾<contrib contrib-type="author">
            <string-name>Berk Hess</string-name>
            <email>hess@kth.se</email>
            <xref ref-type="aff" rid="aff0">0</xref>
            <xref ref-type="aff" rid="aff1">1</xref>
            <xref ref-type="aff" rid="aff2">2</xref>
          </contrib>
```


Figure 5: Metadata title, journal and author in XML extracted by Cermine.


```
▾<kwd-group>
    <kwd>Alchemical free energy calculations </kwd>
    <kwd>Lambda path optimization </kwd>
    <kwd>Accelerated weight histogram method</kwd>
  </kwd-group>
```


Figure 6: Metadata keywords in XML extracted by Cermine.

As can be easily observed in Fig. 5. and fig. 6., Cermine correctly extracts all the required metadata; therefore, all the cells in Fig. 7. are green[10].

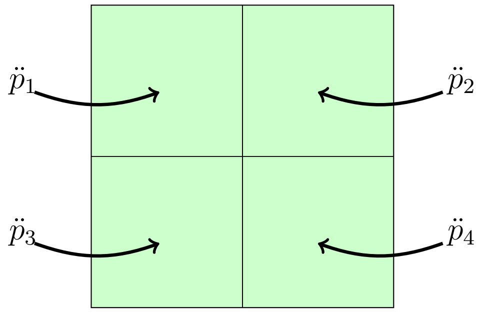

Figure 7: A consistent and complete toy world.

In this case, the metadata extraction performed on the document can be considered both consistent and complete based on the previously defined criteria. Specifically, completeness is achieved by ensuring that the atomic metadata accurately represents all relevant aspects of the document, capturing only the necessary content without omissions or extraneous additions. Consistency is maintained by aligning atomic metadata strictly with the properties of the document, avoiding overlaps between different metadata classes – such as ensuring that abstract metadata remains distinct from keywords metadata. These adapted definitions of completeness and consistency, tailored to the domain of metadata extraction from scientific articles, offer a reliable framework for evaluating the accuracy and quality of the extracted information. Naturally, since the extraction is correct and complete, it is straightforward to verify that specific propositions follow as consequences. As examples, I demonstrate the following logical consequences using the method of *analytic tableaux* [19].

Example 1.

$$(\ddot{p}_1 \to \ddot{p}_2) \land (\ddot{p}_2 \to \ddot{p}_1), \ddot{p}_1 \Vdash \ddot{p}_2 \tag{13}$$

To determine whether $\ddot{p}_2$ is a logical consequence, it is sufficient to construct a countermodel with true premises and a false conclusion.

Proof.

$$(\ddot{p}_1 \to \ddot{p}_2) \land (\ddot{p}_2 \to \ddot{p}_1), \ddot{p}_1 \nVdash \ddot{p}_2$$

[10] In Fig. 7., for the sake of convenience, I report only the green colour of the cell without writing the relevant atomic metadata in each cell.

1.	$(\ddot{p}_1 \rightarrow \ddot{p}_2) \wedge (\ddot{p}_2 \rightarrow \ddot{p}_1)$	Premise
2.	$\ddot{p}_1$	Premise
3.	$\neg\ddot{p}_2$	Denied conclusion
4.	$(\ddot{p}_1 \rightarrow \ddot{p}_2)$	from 1.
5.	$(\ddot{p}_2 \rightarrow \ddot{p}_1)$	from 1.
6.	$\neg\ddot{p}_1 \times \qquad \ddot{p}_2 \times$	from 4.

□

Since the tree is closed ($\times$), there is no countermodel for the sequence; thus, it is never the case that all premises are true and the conclusion is false. Therefore, the sequence is valid and sound, and $\ddot{p}_2$ is a logical consequence[11].

Example 2.

$$\ddot{p}_1 \rightarrow (\ddot{p}_2 \rightarrow \ddot{p}_3), \ddot{p}_1 \rightarrow \neg\ddot{p}_2 \Vdash \ddot{p}_1 \rightarrow \ddot{p}_3 \tag{14}$$

Proof.

$$\ddot{p}_1 \rightarrow (\ddot{p}_2 \rightarrow \ddot{p}_3), \ddot{p}_1 \rightarrow \neg\ddot{p}_2 \nVdash \ddot{p}_1 \rightarrow \ddot{p}_3$$

1.	$\ddot{p}_1 \rightarrow (\ddot{p}_2 \rightarrow \ddot{p}_3)$	Premise
2.	$\ddot{p}_1 \rightarrow \neg\ddot{p}_2$	Premise
3.	$\neg(\ddot{p}_1 \rightarrow \ddot{p}_3)$	Denied conclusion
4.	$\ddot{p}_1$	from 3.
5.	$\neg\ddot{p}_3$	from 3.
6.	$\neg\ddot{p}_1 \times \qquad \ddot{p}_2 \rightarrow \ddot{p}_3$	from 1.
7.	$\neg\ddot{p}_1 \times \qquad \neg\ddot{p}_2 \times$	from 2.
8.	$\neg\ddot{p}_2 \odot \qquad \ddot{p}_3 \times$	from 6.

□

Since the tree is not closed ($\odot$), $\ddot{p}_1 \rightarrow \ddot{p}_3$ is not a logical consequence.

5 Conclusion

In this paper, I have examined how foundational principles of modal and algebraic logic can meaningfully address the critical issue of data veracity within descriptive metadata for scientific articles. By aligning the zones of an article with an

[11]For an introduction to *proof theory* see [20], and for its application to data science see [21].

algebraic situation space, I have crafted a framework in which atomic metadata functions as verifiable, representative components of the document world. This approach facilitates a rigorous assessment of metadata veracity by grounding it in well-defined logical standards of completeness and consistency: completeness ensures that all relevant facets of an article are represented fully, while consistency confirms that the metadata accurately reflects the document's structure without contradictions. The significance of data veracity within the broader context of data science and Big Data cannot be overstated. With the exponential growth of digital data, the need for reliable and quality information has become paramount. By applying situation-based modal logic to metadata veracity, this study not only advances a logical, structured methodology for verifying the accuracy of extracted metadata but also emphasizes that quality in Big Data depends heavily on data veracity. High-quality data infrastructures – especially in scientific domains – rely on rigorous, consistent, and comprehensive data, which is vital for supporting valid analysis and trustworthy insights in fields ranging from academia to industry. This work represents an important step toward reinforcing the integrity of scientific metadata in a digital age, where vast amounts of data are at risk of being compromised by inaccuracies or inconsistencies. The future developments of my research will focus on: *i*) refining the logical tools employed by applying mathematical methods characteristic of topology [22]; and *ii*) implementing an innovative framework of knowledge extraction based on mathematical logic.

6 Acknowledgments

This paper would not have been possible without a solid foundation in logic, data science, and digital humanities. In logic, I am deeply grateful to Sergio Galvan and Alessandro Giordani for their invaluable guidance and teachings during my years at the *Università Cattolica del Sacro Cuore* in Milan, which played a pivotal role in my academic development. In the fields of data science and digital humanities, I am particularly thankful to Antonietta Folino and Anna Rovella from the *Università della Calabria*, as well as to Elena Cardillo and Maria Taverniti of the *Knowledge Organization Systems* research group at the *Istituto di Informatica e Telematica* (IIT) of the CNR, for their important support.

References

[1] L. Berti-Équille and J. Borge-Holthoefer. Introduction to Data Veracity. In: *Veracity of Data. Synthesis Lectures on Data Management*. Springer, 2016.

[2] C. Batini, A. Rula, M. Scannapieco and G. Viscusi. From data quality to big data quality. *Journal Of Database Management*, 26 (1), pp. 60–82, 2015.

[3] C. Batini and M. Scannapieca. *Data Quality. Concepts, Methodologies and Techniques*. Springer, Berlin, 2006.

[4] S. Cuconato. A Four-Valued Epistemic Logic for Metadata Modelling from Medical Articles on Pain Therapies. In: Das, A.K., Nayak, J., Naik, B., Vimal, S., Pelusi, D. (eds.) *Computational Intelligence in Pattern Recognition*. CIPR 2023. Lecture Notes in Networks and Systems, vol 725, pp. 631–640. Springer, Singapore, 2023.

[5] W.V. Quine. *Ontological relativity and other essays*. Columbia University Press, New York, 1969.

[6] D. Tkaczyk, P. Szostek, P. Jan Dendek, M. Fedoryszak, Ł Bolikowski. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition*, 18(4), pp. 317–35, 2015.

[7] S. Cuconato. Modal Logic, Probability and Machine Learning Systems for Metadata Extraction. *Ratio Mathematica*, 53, pp. 7–24, 2024.

[8] S. Cuconato. A Logical Language for Reasoning about Democratic Decision–Making. *Proceedings of 1th International Conference on Advanced Network Technologies and Computational Intelligence* (ICANTCI 2024). In: Singh, J., Goyal, S.B., Kumar, M., Mittall, R. (eds.) Advanced Network Technologies and Computational Intelligence. ICANTCI 2024. Communications in Computer and Information Science, vol 2382, pp. 293–302. Springer, Cham, 2025.

[9] A. Giordani. Situation-Based Connexive Logic. *Studia Logica*, 112, pp. 295–323, 2024.

[10] J. Austin. Truth. *Proceedings of the Aristotelian Society*, Supplementary Volume, 24, pp. 111–128, 1950.

[11] K. Fine. A theory of truth-maker content I: Conjunction, disjunction and negation. *Journal of Philosophical Logic*, 46, pp. 625–674, 2017.

[12] K. Fine. A theory of truth-maker content II: Subject-matter, common content, remainder and ground. *Journal of Philosophical Logic*, 46, pp. 675–702, 2017.

[13] J. Barwise. *The Situation in Logic*. Center for the Study of Language (CSLI), Stanford, 1989.

[14] G. Grätzer. *General Lattice Theory*. Birkhäuser Basel, Springer, 1978.

[15] M. Lundborg, J. Lidmar and B. Hess. On the Path to Optimal Alchemistry. *The Protein Journal*, 42, pp. 477–489, 2023.

[16] F. Berto. *Existence as a Real Property. The Ontology of Meinongianism.* Springer, 2013.

[17] S. Cuconato. Can Nonexistent Mathematical Objects Make a Difference? Meinongianism, Indispensability Argument and Mathematical Entanglement. *Philosophical Inquiries*, 12(1), pp. 61–77, 2024.

[18] S. Cuconato. Oggetti Matematici Non-Esistenti come Truthmakers: Meinonghianismo e l'Argomento di Indispensabilità. *Rivista di Estetica*, 85 (1/24), pp. 168–183, 2024.

[19] A. Das, S. Negri. *Automated Reasoning with Analytic Tableaux and Related Methods*. 30th International Conference, TABLEAUX 2021, Birmingham, UK, 2021.

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

[21] S. Cuconato. Jaśkowski–Fitch Natural Deduction Proofs for Metadata Extraction through Machine Learning Systems. *Proceedings of 6th International Conference on Computation Intelligence in Pattern Recognition*. CIPR 2024. In: Das, A.K., Nayak, J., Naik, B., Vimal, S., Pelusi, D. (eds.) *Computational Intelligence in Pattern Recognition. Lecture Notes in Networks and Systems*, Springer, 2025. (*forthcoming*)

[22] M. Aiello, I. Pratt-Hartmann and J. van Benthem (eds.). *Handbook of Spatial Logics*. Springer, 2007.