New Terminological Approaches for New Heritages and Corpora: The ITinHeritage Project

Caroline Djambian, Micaela Rossi, Giada d'Ippolito, Emrick Poncet, Pierre Maret

New Terminological Approaches for New Heritages and Corpora: The ITinHeritage Project

Caroline Djambian[1], Micaela Rossi[2], Giada D'Ippolito[3], Emrick Poncet[4] and Pierre Maret[5]

[1]*Grenoble Alpes University, Lab GRESEC (Groupe de recherche sur les enjeux de la communication)*
[2]*Genoa University, Dipartimento di Lingue e culture moderne*
[3]*Genoa University, Dipartimento di Lingue e culture moderne*
[4]*Saint Etienne University and Grenoble Alpes Lab GRESEC (Groupe de recherche sur les enjeux de la communication)*
[5]*Saint Etienne University, Lab H. Curien*

Abstract

Safeguarding Information Technology (IT) heritage is a matter of general interest. This is the aim of the ITinHeritage project, which takes a heritage-based approach to IT museums in France and around the world. The unprecedented study of this heritage leads us to understand how to perpetuate and mediate contemporary scientific and technical knowledge, of which data is the new heritage. They also constitute the new corpora, which means that terminology work needs to be rethought. The ITinHeritage project aims to develop an innovative approach to digital humanities by creating a corpus from the collections of IT museums, organised in the form of a knowledge graph, and by using methods and tools that combine traditional linguistic approaches and explorations in the computer sciences. Big data, AI and NLP are thus being used to explore, enhance and perpetuate their own heritage.

Keywords

Terminology, Multilingualism, Lexical extraction, Automatic Language Processing (ALP), Artificial Intelligence (AI), Knowledge graph, Ontology, Linked Open Data

1. An evolving subject for study: Information Technology (IT) heritage

Technoscience, the science embodied in technology, not only makes the world more intelligible, but it also transforms and impacts the world in unprecedented proportions and at unprecedented speed. Technoscience is the foundation of contemporary works of humanity. Information Technologies (IT), defined by UNESCO (2023) as "the set of tools and technological resources that enable information to be transmitted, recorded, created, shared or exchanged", are the emblem of these technologies and their specific features. But to this day, even though IT permeates and shapes our societies, this field remains poorly defined, because the rapid and massive evolution has left little time for study and heritage development.

3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2024, June 27-28, 2024, Granada, Spain.
caroline.djambian@univ-grenoble-alpes.fr (C. Djambian); micaela.rossi@unige.it (M. Rossi); giadadippolito30@gmail.com (G. D'Ippolito); emrick.poncet@univ-grenoble-alpes.fr (E. Poncet); pierre.maret@univ-st-etienne.fr (P. Maret)

Looking at this heritage is a crucial part of understanding today's world and how society has evolved since the second half of the twentieth century. But how can we define it, represent it, promote it and preserve it in its rapid and proliferating evolution? Our approach, which is resolutely heritage-based and interdisciplinary, takes as its starting point the IT museum spaces that have initiated this heritage enhancement. The Musée des Arts et Métiers (MAM, France), ACONIT (Grenoble, France), the London Science Museum (London, UK), the NAM-IP (Namur, Belgium), the Museo degli strumenti per il calcolo (Pisa, Italy), the Home Computer Museum (Helmond, Netherlands) and the HNF (Paderborn, Germany) are actively participating in our expanding network. The ITinHeritage research project thus aims to conduct an epistemological reflection on the issue of IT by placing at the heart of the question of heritage the history of a societal mutation and the emergence of new knowledge in terms of representations of the world.

Our starting point is the as yet unexplored definition of this labile heritage, whose knowledge is crystallised in various forms. Firstly, through the expression of its explicit knowledge, through its physical objects and its digital objects: the digitised artefact and its documentation, the software which represents 1/3 of certain collections and above all the data which forms the new collections of science. Secondly, through the expression of the tacit knowledge of IT, which has not yet been questioned or given a heritage status. This tacit knowledge is expressed at the experiential level. Whereas 'pure knowledge' [1] is 'exoteric' [2], i.e. concretised in writing or in an object, 'empirical knowledge', the emperia or métis (Aristotle), can only be expressed by its bearer in action. Gathering and transmitting this tacit, "esoteric" knowledge is a real challenge because, to be captured, emperia needs to be shaped into gestures and discourse. The language of specialisation is the main expression of this knowledge, i.e. of experience of the realities of the world [3].

We are therefore placing language at the heart of the ITinHeritage project by combining complementary and innovative approaches aimed at the interaction between terminology and ontology. This link is not a new theme. From this theoretical perspective, Sager (1990) adopts an onomasiological approach, but based on corpora of lexical data collections, adapted to the creation of special linguistic vocabularies. The descriptive/socio-terminological theory in France, theorized by Gaudin [4] and Boulanger [5], leads to the study of the actual use of language, the importance of external influences, and the diachronic evolution of language. This last stage was subsequently given prominence in cognitive theory, which studies the cognitive processes involved in the use and processing of language. Its main exponent is Geeraerts. In our case, we are projecting ourselves into the work of Temmerman (2000) and her formulation of the unit of understanding (UoU), i.e. categories whose prototypical structures are constantly evolving, which requires us to take into account the conceptual nature of terminologies, as well as into the wake of the research of Christophe Roche and his onto-terminological approach[6].

2. Information Technologies (IT) corpora

Data are not only the new objects (collections) of heritage, they are also the new corpus. To build our corpus, we chose to model the metadata of museum collections in the form of a knowledge graph, in the Semantic Web paradigm, which promotes links between data and is

easily augmented by the arrival of new data. A crucible of IT heritage, our Knowledge Graph was developed using the Wikibase environment, after collecting and harmonizing the museums' metadata (native XML, PDF, etc. formats, free fields) and then converting it to CSV, RDF and FAIR standards so that it could be opened up and linked on the web (Linked Open Data, LOD). The Knowledge Graph is structured by a first-level ontology based on CIDOC-CRM (International Committee for Documentation of the International Council of Museums, Conceptual Reference Model) and the EDM (Europeana Data Model) on the Protégé environment.

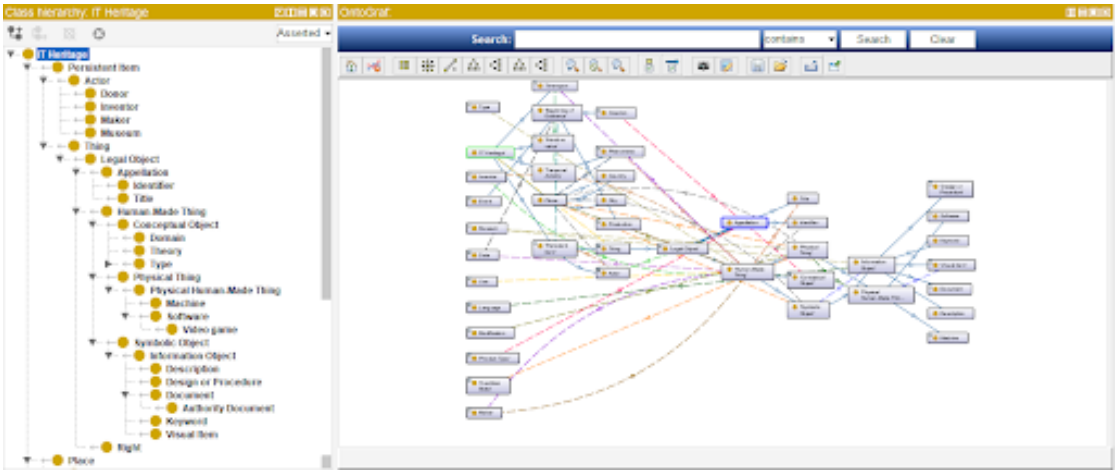

Figure 1: Extract from the first level ontology that structures the ITinHeritage knowledge graph:

The Knowledge Graph currently contains more than 25,500 artifacts and is set to grow. It perpetuates and opens up the IT heritage, allowing it to be exploited, and above all it offers us a choice textual corpus, very representative of the new corpora that terminologists are facing today, resulting from Big Data. This multilingual corpus (French, English, German, and Italian) is the basis for a study of the discourses that represent IT knowledge, which we will describe in detail. With the help of experts in the field (Pr Marie Gevers from NAM-IP and Xavier Heron from ACONIT), we are embarking an onto-terminological work through the construction of a quadrilingual (FR, EN, IT, GE – by now, the corpus has been explored for French and English) and multimedia dictionary, and an ontology of IT field, based upon the founding models of terminological work (ISO standards 5078, 704 and 1087).

3. From terms to concepts in the Information Technology (IT) field: the appropriation of scientific and technical knowledge versus technicality and variety of languages?

If we focus on language, we must take it in all its complexity. It is true that, as Bertrand Russell notes, in an ideal language there should be only one word for a single object, and any complex object would be expressed by a combination of words, each for each characteristic of the object [7]. But Democritus already noted in antiquity that: 1) different objects are often designated

by the same name; 2) the same object is often designated by different names; 3) the names designating an object may vary over time; 4) the reasons for which names are linked to objects vary greatly [8]. Thus, the same object is designated by different terms in different languages. However, "there are scientific and technical fields that require a conceptualization of the world and the creation of unambiguous names for its components" [9]. This is mandatory for making knowledge accessible, especially when we are addressing a wide audience.

A fortiori, when aiming to build a science dictionary, the dissemination of scientific and technical knowledge requires explaining the significance of the names of concepts in a field, i.e. specialist terms. The outreach and didactic purpose, places greater emphasis on the notional dimension, its accuracy, and its deciphering to cover all the knowledge in a field. In essence, they are aimed at a more restricted audience than encyclopedic dictionaries, even if the primary intention is to disseminate knowledge as widely as possible and reach everyone. Their construction involves, in our case, the participation of historians of science and especially of computing. The transmission of scientific and technical concepts to as many people as possible, necessarily requires the integration of a meta-linguistic discourse and textual and semantic strategies, to translate the accumulation of specialized terms into an understandable language.

In terminology, a descriptive strategy will focus on differentiating notions from linguistic links, to make the names of concepts, that are opaque to the uninitiated, or even unknown, readable, and appropriable. The aim here is to bring scientific concepts into the cultural mainstream [10]. This is because the materiality of language is just as much an obstacle to access to specialized knowledge, as the lack of mediation of scientific and technical objects. The linguistic forms (signifiers) of science can turn out to be just as unintelligible as the object they name. Paraphrasing and syntax are two tools that can avoid unwelcome heaviness. But in the transmission of knowledge, it is the concept that needs to be integrated rather than its name, even if Putnam [11] shows that we can talk about knowledge without having acquired it. Thus, it is not enough to acquire a lexical system; we need to acquire a notional system. However, a new word is apprehended by attaching it to and differentiating it from a network of units already held. But how do we go about constructing meaning when these pre-requisite units do not yet exist? There are two possibilities: the exploitation of semantic-syntactic relations and lexical relations. Semantic-syntactic relations such as "typical object", "typical action" and "typical agent" (Lerat, 1987 and 1988), to which Gaudin (1995) adds "typical application", are useful for transmitting complex knowledge. They list the typical collocations of a lexical unit to give, through this combination, a precise idea of the use made of the named object. Here is an example of metadata from inventory collections for the artifact « Tabulator BS 120 » from ACONIT Conservatory (Grenoble, France):

- «La tabulatrice Bull BS 120 se compose d'un lecteur de cartes (traitant 150 cartes par minutes), d'une imprimante (cadence de 150 lignes par minute sur une largeur de 92 colonnes), d'une perforatrice de cartes (débitant soixante-quinze cartes par minute). Elle dispose d'un calculateur mécanique qui lui permet d'exécuter les quatre opérations arithmétiques, des opérations logiques et de mémoriser des informations. Elle offre un système de programmation amovible : le "tableau de connexion", qui était spécialement câblé pour chaque traitement. Sa technologie est entièrement électromécanique.».
- **Typical action:** lire, additionner et imprimer

- **Typical object:** caractères
- **Typical agent:** calculateur mécanique, lecteur de cartes, imprimante, perforatrice de cartes (ateliers mécanographiques à cartes perforées)
- **Typical application:** calcul électromécanique

Particularly well-suited to popularizing science, semantic-syntactic relations lead the work of textual analysis toward the modeling of knowledge. A fortiori, they make it possible to express the practices surrounding the object. Putting its application into context can only be conducive to understanding, even if this description is not sufficient for the acquisition of knowledge. The path towards the concept to be understood is thus supported by a description of the object's properties, i.e. a "making sense" and a delimitation, with these properties echoing the characteristics of the concept.

But the scale of complexity of a specialist language for the layman varies enormously from one field to another. In the example cited above, the semantic-syntactic relationships presented are not enough to provide access to meaning. Specialized terms need to be translated, otherwise, there is a risk that too many obscure terms will be brought together and accumulate, creating a barrier for the uninitiated. For example, the typical application "electromechanical calculation" can be presented as a "precursor of computer programming", which provides the reader with a reference point. Thus, not all scientific languages are equal in difficulty in understanding their concepts, and they do not have the same inclusions in everyday language. Specialized language is not a whole. So many obstacles stand in the way of understanding by the layman: complex terms (mechanographic workshops with punched cards), eponyms (Turing machine, Moore's law), words' morphologies such as acronyms, complicated by version numbers (IBM 1130), special characters, etc. However, these signifiers can also serve as echoes of fuzzy knowledge, a visual and semantic anchor of recognition for the reader. Contemporary science is characterized not only by its highly technical nature but also by its social roots. Technical objects are at the heart of our daily lives, all the more so for science such as Information Technologies. Many terms are an integral part of our knowledge, such as a keyboard or computer mouse, and it is through their typical application that meaning can be built up towards more complex notions and names attached to these objects. Terminological complexity is thus found between the common language and the specialist language. But language uses can also differ within the same specialty, geographically (diatopy), over time according to technological developments (diachrony), and between communities of practice (diastracy). For example, a "first generation calculator" for the computer amateur (source: Wikipedia), will be referred to as a "mechanical calculator" by a computing history expert, or as an "ancient calculator" by a general history expert. For the initiated, one of the objects embodying this concept is an "abacus" ("abaque" in French), or "first digital tablet". For the general public of the 21st century, however, the concept has a completely different connotation, and the term is associated with a modern digital pad (portable computer), whereas the object, which dates back to ancient times (it has been found as far as 500 BC), is actually made up of clay balls and tokens, the "calculi" (lat.). "These were used for arithmetic until 7000 BC, and evolved into a device with rows of moving parts, better known from 500 BC under the name of "abacus" ("boulier" in French, which distinguishes the two types of objects in their diachronic evolution, contrary to English). Differences in language, particularly among experts, stem from different representations of the world and different

strategies: "... we only manipulate reality through the representations we have of it" [12]. In the field of Information Technologies, we can therefore find very divergent denominations: those of the general public, the amateur, the expert in the world of technology or heritage, etc. Museum inventories, for example, use highly controlled language, demonstrating, just as much as the accumulation of knowledge around artifacts with documentation, a mastery of the subject, setting up curators as credible guardians of an almost sacralized heritage. The "Baby" for the amateur is called by museums "Manchester Baby" because of where it was built, or "Manchester Small-Scale Experimental Machine" (London Science Museum) in reference to its history: the "Small-Scale Experimental Machine" (SSEM) was the world's first von Neumann architecture machine. Built at Manchester's Victoria University by Frederic Calland Williams, Tom Kilburn, and Geoff Tootill in 1948, it is what is known to the general public as the first "computer". According to an expert in the history of computing, mainstream "computers" are "stored-program computers, in the strict sense of the term, i.e. von Neumann-type computers, which have a very precise meaning: they are electronic calculating machines with a central memory large enough to hold the program being executed, as well as the data". We are therefore faced with highly diversified linguistic uses, which can only lead to a blurring of concepts and names. On the one hand, there is the everyday language, and on the other, there are the specialist languages, which do not have the same status and are based on extralinguistic realities shared by sub-communities in the same field. In these languages, words can have a different linguistic weight, for example, when they are themselves the name of a concept [13]. In this inclusive relationship between everyday language and its specialized subsets, words are the only tangible elements available to us to capture the representations to which they refer. We naturally find them in texts that can serve as a basis for terminology work. Linguistic analysis of corpora enables us to extract syntagms that can be linked together by lexical relations. Lexical gender and partitive relationships are essential to the acquisition of new knowledge, and ISO 704 identifies them as the foundation of terminology work. Hyperonymy, hyponymy, meronymy, antonymy, or isonymy, effectively situate the new term in a notional context. Hyperonymy in a genus/species relationship, hyponymy in class descriptions, antonymy, and especially isonymy, in notional differentiation. "To determine the meaning of a unit, the reader needs contrasts, and the canonical definition provides only one mode of category construction, that of specialization with the genus, defining it" [10]. Isonymy, as "any relationship linking two competing units, usually at the same level, without it being possible to establish a hierarchy that is valid from all points of view" [14], makes it possible to establish a fine-grained semantic relationship between closely related concepts, without any hierarchical link, and situating them as part of a whole. Concepts are thus described in a grouped manner. As we shall see, these lexical relationships are highly represented in our corpus, which is drawn from science museum collections, according to museum mediation strategies that are pedagogical (educating the public), narrative (telling the story of a technology and its inventors or producers) or descriptive (describing an object according to its components or functions).

4. Information Technologies (IT) lexicons

The study of computer terminology is undoubtedly one of the fields that has most interested terminologists, who have delved into its morphological and semantic aspects (among others, see [15]), its uses in specialized discourse and its textual dimension, with particular reference to corpus linguistics (for an initial fondamental study, see Condamines, 2005), and issues related to interlingual comparison (among others, [16]). Particularly interesting in our perspective is L'Homme, 2008[17], comparing terminological, ontological and general resources in describing computing terminology[1]. The ITinHeritage project is heritage-based, deeply interdisciplinary in nature, and the dialogue between disciplines presupposes an interrogation of fundamental notions such as term and concept from different points of view - namely, from a knowledge-based approach and from a linguistic-based one. As L'Homme states, "Given the differences between the assumption of lexical-based and knowledge-based approaches and the principle on which they rely, the question is whether they can be used simultaneously in terminology work" [18]. In this perspective, by its focus on heritage-based, interdisciplinary vision, the ITinHeritage project can be considered as an attempt to find an answer to this question. The main issue that interests us at this stage of the project is the identification of terminological units and lexical relations, necessary to structure domain knowledge ; thus, our analysis stands in continuity with many studies in the field of terminology, which question the interface between terminology and knowledge-based approaches [19][20][21][22].

In a first phase, we postulate that the study of the linguistic uses [23] in the IT heritage field, will enable us to identify names of concepts and their significations, socially stabilized within the various communities that found this domain, by identifying the relationships mentioned above. This detailed analysis will ultimately enable us to construct a dialectic between experts, science, and the public, and to observe the joint evolution of language and technology [24]. As the methodology in textual terminology is based on the triptych specialized corpus - experts - digital tools [25], our terminological work begins with a semasiological approach, through lexical extraction from museum metadata and linguistic analysis of our corpus. The corpus, composed by the collection of metadata from museums involved in the project, is built following the text selection criteria ISO/DIS 5078:2023. It is precisely from these metadata that a first terminological extraction, always based on the regulatory criteria of ISO 5078, is currently carried out. Our extraction approach, which will be applied to all languages included in the project, can be considered semi-automatic, comprising an initial automatic selection by extraction tools (Termostat and Sketch Engine) and a subsequent manual intervention for lexical cleaning. We have relied on a predominantly hybrid approach, combining statistical techniques initially and then linguistic ones. We will have the opportunity to analyze these techniques in detail with respective examples. Starting with the French corpus, whose data come from the ACONIT museum, the following fields were taken into consideration: description, name, model, and use. In this corpus, Description and Use are the only two fields composed of multiple textual

[1] L'Homme (2008), made a lexical extraction work regarding the computing field. She extracted the 75 most frequent candidate terms from a corpus of specialized resources in the field of computer science using TermoStat. She demonstrated how, in terms of formal coverage, the resource that exhibited a greater presence of these terms, and therefore more completeness, is WordNet, a general-purpose lexical database, compared to domain-specific online dictionaries and ontologies. These latter resources even yielded differing results among themselves.

sentences, while name and model can be defined as made up by a sort of controlled vocabulary.

For the first two fields, it was necessary to rely on a strategy of automatic terminological extraction. Various tools were tested and compared, including Sketch Engine [26], TermoStat [27], and we also relied on the NLTK package of the Python programming language [28] as far as programming languages are concerned, they are now considered relevant alternatives to traditional extraction tools when facing corpora derived from big data [29]. However, after a first attempt, in the initial extraction phase done by Termostat as well as with the NLKT library, we encountered a high level of noise, which would have required an excessively demanding manual intervention. Finally, Sketch Engine, a tool for textual management and analysis for extraction, proved to be the best option. The comparison corpus is the French Web Corpus 2023 (frTenTen23), composed of texts collected from the internet, totaling approximately 24 billion words in the French language. The corpus annotates words with POS tags using the FreeLing tool. This comparison corpus belongs to the TenTen corpus family, extending across 40 different languages (Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V., 2013). Thanks to the TenTen corpora, we can use the same procedure applied to the French corpus for other languages as well. For the terminological extraction by the Sketch Engine function keywords, the selection filter was set to a maximum of 5000 items, considering with particular attention the MULTI-WORD TERMS results. The first SINGLE-WORD terms obtained, such as micro-ordinateur, disquette, macintosh, microprocesseur, azerty, hewlett-packard, powerbook, olivetti, alphanumérique can in many cases be considered quite generic. We decided to focus on multi-word terms which seemed to be the most frequent terminological pattern in IT. Regarding linguistic analysis of the extracted terms, the TermoStat software was initially preferred: simple and complex terms were extracted by selecting lexical categories such as nouns, adjectives, and verbs. TermoStat uses a method of contrast between specialized and non-specialized corpora to identify terms. Those of our greater interest are terms with the highest specificity score, with a lower frequency in the general reference corpus, but more representative in the specialized domain. TermoStat has some limitations: it is not possible, for example, to extract acronyms or terms beginning with an uppercase letter followed by a series of numbers. Extraction tools tend to overlook these patterns, while they represent a major terminological source in the IT industry. In most cases, these are names of machines and their models or manufacturers - e.g., Gamma 30, IBM 1130, ...) and in this case, regular expressions and the Corpus Query Language (CQL) of the Concordance section of Sketch Engine can be used, allowing the analysis of words, sentences, documents in various possible contexts, extracting and analyzing more complex grammatical patterns. However, a first analysis provide some useful insight about the morphological patterns of candidate terms, as shown below:

Some preliminary observations can be made, on the basis of this first linguistic analysis. The description and the use fields present different characteristics as for the identified candidate terms. The description field focuses on machines and their components, as shown by the 10 most frequent candidate terms (classified on the basis of their specificity score):

Verbs, which make up an interesting percentage of this sub-corpus, are very often markers of meronymic relations [30], as in the case of posséder (132 occurrences), contenir (101 occurrences), comprendre (90 occurrences), comporter (30 occurrences). The markers often indicate references to the composition of the machines or their technical functioning. In the case of adjectives, they mostly refer to the shape or functions of the machines. For example, 'clavier,' the most frequent

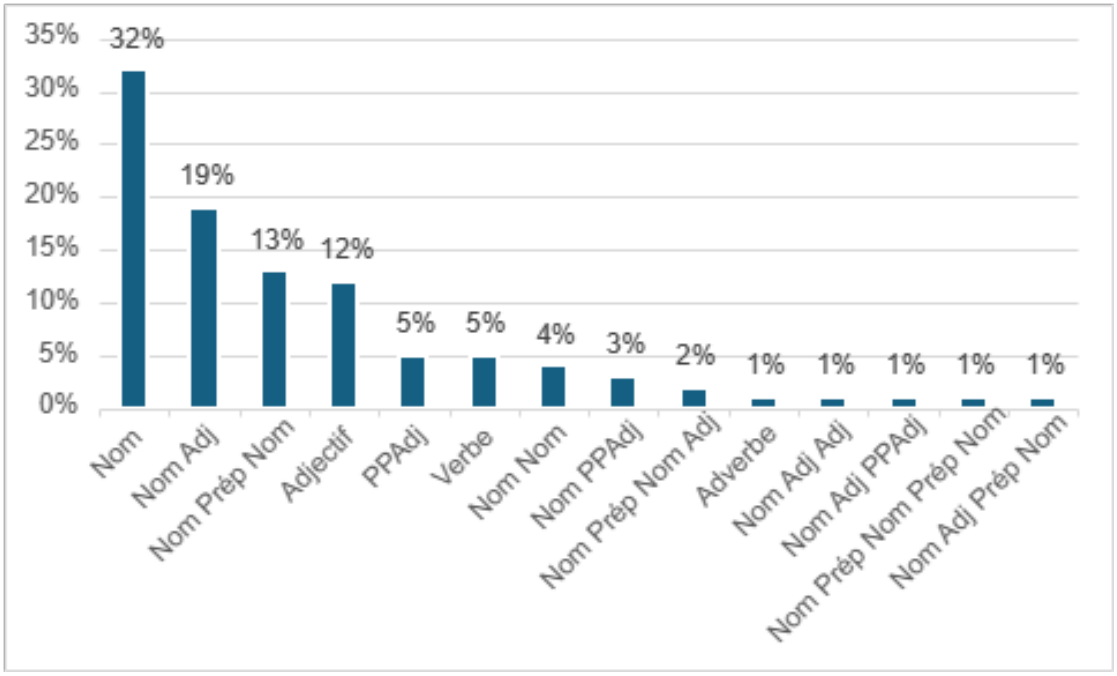

Figure 2: ACONIT corpus - description field - 2998 candidate terms:

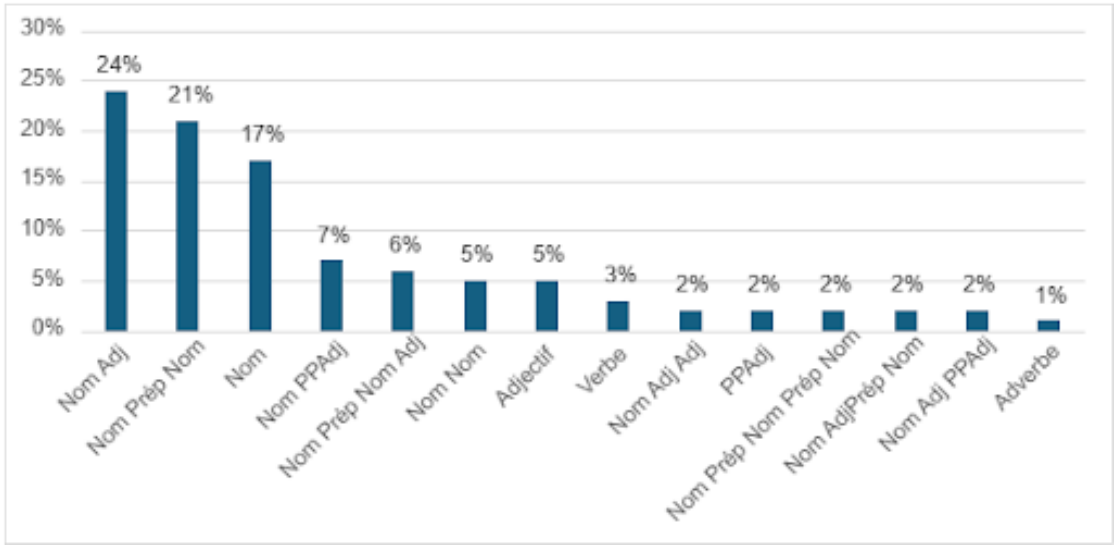

Figure 3: ACONIT corpus - use field - 5967 candidate terms:

noun among the candidates provided by TermoStat (Frequency 509, Score (Specificity) 269.89), is associated with adjectives such as alphanumérique (153.97), numérique (44.83). Whereas the Description field involves descriptive museum mediation centered on the components of the object, the Use field involves narrative mediation strategy. It reveals a more complex

Table 1
Frequency and Specificity of Regrouping Candidates

Group candidate	Frequency	Specificity
Keyboard	509	269.89
Floppy disk	296	231.41
Connector	258	227.80
Microcomputer	244	214.53
	212	202.31
Button	242	177.87
Floppy disk drive	150	174.22
Key	325	170.24
Printer	137	154.90
Machine	443	145.76

structure, where the terminological extraction focuses not only on artifacts, but also on practices and actors involved in IT history, as suggested by the presence of the verb utiliser and of the nouns utilisateur et utilization. Verbs in this section are more often focused on relating specific actions or processes (permettre 471 occurrences, pouvoir 308 occurrences, servir 123 occurrences, agir 72 occurrences, fonctionner 60 occurrences), but we can also find verbal conceptual relation markers such as appartenir (9 occurrences), whose related terms highlight meronymic relationships: appartenir-lignée (73.17); appartenir-famille (56.93); appartenir-micro-ordinateur (48.81). Moving to the English language corpus, made up of metadata provided by the London Science Museum, the fields for lexical extraction taken into consideration are the same as for the French corpus in order to obtain a reliable comparative study. Now we can make an initial comparison between corpora in different languages. For example, the results of the Description field extraction show that this field most focuses on artifacts and concrete objects contextualized within a history or practice, and in this respect is similar to the mediation strategy observed in the Use field of the French corpus. Moving to another museum and another culture changes the relationship with the object and the representations of the world to which it belongs. Here, the morphological pattern Adj+Noun is more productive in this field, with examples such as personal computer (110 occurrences), video game (84 occurrences), electronic calculator (53 occurrences), electronic component (41 occurrences). The most frequent verbs focus on concrete actions: manufacture (195 occurrences), build (120 occurrences), work (101 occurrences), even if we can also find an important number of occurrences for verbs such as include (51 occurrences), which can be considered as markers of meronymic relations.

5. The need for more in-depth approaches

This semasiological approach results in lexicons that reflect the initial corpus and highlight rather than resolve the linguistic variety of a domain. It is not yet a consensual conceptualization. But if we are interested in the way experts name their domain, we are interested in the way they conceptualize it. However, our experience has shown us how difficult it is to draw up a semantic network solely on the basis of this semasiological work [31]. It poses the problem

of finding only the variety of names of domain concepts in the texts, and not the concepts themselves. This type of study, although essential for laying the foundations of terminology work, stops at the meaning of words observed in discourse, defined according to their uses. To compensate for linguistic variations and bring out a common meaning, it is advisable to take an interest in the extralinguistic part of terminology work, centered on the relationship of the concept to the object. The meaning of a word is intended to be independent of its uses and is defined as a meaning that is consensually standardized within a community, referring to a conceptualization of the world. The meaning of a word is "signification actualized in discourse" [12]. Thus, "it should be remembered that all terminological work should be based on concepts and not on terms" (Felber, 1984). Terms are of interest to the terminologist because they denote a concept, a kind of bijective form, a bridge, reflecting the continuous dynamic between the linguistic part of the real world and the extralinguistic part of the symbolic one. The notional (or conceptual) world is the translation of how we apprehend objects in the real world. The standardization required by terminology work makes it possible to fix this consensual notional system within a community of practice. In order to take into account the specificities and variations of the linguistic system, the notional system is more easily constructed using an onomasiological approach. In this formalization based on the convention in the Latin sense of foedus, the names of concepts must be situated within the notional system as a whole, even if this means artificializing them to provide a signification that goes beyond usage. Structuring concepts according to subsumption and difference relations enables us to build a standardized vocabulary that blurs the varieties and ambiguities of natural language. The linguistic relations mentioned above and observed in our corpus, above all, the genus and the merological relation, enable a detailed description of objects in that they provide the notional environment (specific features or constituents of an object). In this way, ""ontological relations" are defined as "indirect relationships between the notions", the most important of which is the merological (partitive) relationship" [12]. So, objects are defined in the sensible or intellectual reality, according to the properties that the experience of empirical practice forms for them in the real world, to which the linguistic stratum corresponds. By abstraction, these notions (concepts) reflect and are organized according to these properties in the meta-linguistic stratum (the symbolic world), which are translated into a set of characteristics derived from the reason. The organization of concepts to each other within the notional system therefore requires us to question the very essence of things, their eidetic characteristics.

At this point, the help of experts in the field is mandatory. They have a socio-linguistic responsibility in the transfer of knowledge because they are the only ones to master the discourse of the field, its conceptual representation, and its consensus [11]. Our work therefore is now to model this knowledge using an ontology that reflects the conceptualization of the IT world. The construction of the IT domain's onto-terminology is facilitated by the use of both the Protégé environment for the structuration of the knowledge graph, and the TEDI AI tool [6], which is more accomplished than Protégé for terminology work according to the ISO 704 and 1087 standards [6]. At the same time as building the dictionary, it allows us to define the ontology and analyze the conceptualization phenomena. This novel modeling of IT heritage knowledge based on CIDOC-CRM upper-ontology, will integrate the OWLTime model [32] to be diachronic and represent technological developments in the field. It will complement the first-level ontology by integration via CIDOC classes. Our onto-terminology has to be evolutionarily because IT

cannot be represented as fixed in time when it is constantly being created. For this purpose, we mobilize automatic language processing (ALP), clustering and machine learning to create an automated system for extending the terminology and ontology initially created by linguists. The work is based on data sets recognized by experts. New classes can be created by working on partitive relations. We add a layer of description logic to formally specify the definition of classes using the Protégé environment. Our contributions here focus on the application of automated methods for completing manually constructed onto-terminologies to keep pace with the evolution of the domain and the naming of new objects and knowledge.

In this way, our terminological work is supported by the Computer Science Ontology Classifier (CSO) [33], a tool originally developed to automatically categorize Computer Sciences research papers, based on abstracts, into a comprehensive semantical network, especially the gender and partitive relationship and which we are founding our terminological work. We decided to adapt this advanced AI-driven tool to our specific research needs, employing it to analyze particular descriptive fields, first within our dataset, secondly on web datasets, such as Wiktionary and BDPedia. It will aim us to detect neo-terms relating to new objects in the IT field and allow us to fit these neo-terms into our ontology. This adaptation involves a dive into the classifier's operational mechanics, necessitating modifications to accommodate our unique ontology, which diverges from the broad Computer Sciences focus of the CSO. A key advantage in this adaptation process is the close alignment between our research domain and the original CSO corpus, allowing us to utilize the existing word to vec embedding model [34] which was trained on 4.6 millions English papers in the field of Computer Science and align to the corpus we are using. We remain open to exploring the potential benefits of retraining these embeddings in the future, to see if such refinements could enhance the classifier's accuracy and sensitivity to the nuances of our domain.

Finally, because our work would be useless if inaccessible to the public, we are building a web portal to provide access to the Knowledge Graph. Other platforms federate museum collections [35][36][37] or safeguard software [38], but these projects enhance objects, whereas we aim to enhance knowledge. The platform, designed as a place for transferring knowledge, will offer: navigation via a time map to represent the technological evolution across history; a graph based upon the domain ontology to provide input to the knowledge graph via concepts representing domain knowledge; knowledge completion via the multilingual dictionary; and natural language querying. Querying the Knowledge Graph (RDF) in natural language requires SPARQL (SPARQL Protocol and RDF Query Language) queries. This approach significantly lowers the barrier to entry for users unfamiliar with technical query languages, making the data within our Knowledge Graph accessible to a broader audience. So, here again, we must mobilize NLP, through question-answering (QA) technologies, with IA/IHM tools. SparNatural is a well-known open-source tool available for querying by non-experts [39], which can be used to navigate our knowledge graph based on the domain ontology. Comparatively, the QAnswer tool [40] is end-user-oriented and automatically translates natural language queries into SPARQL. It is intuitive and also learns from user feedback. We are currently assessing the suitability of these two tools for our needs.

6. Conclusion

The Information Technologies heritage is defined by the intrinsic tension between ”objects - languages - representations” [41]. Technologies, terminologies, and conceptualizations in the field are constantly evolving and interrelated. Our work aims to highlight and analyze these developments. It is the starting point for a better understanding of this recent and proliferating field, and of its heritage and dissemination, which are of prime importance, as it bears witness to our contemporary era and its digital and societal transformation. In this sense, the ITinHeritage research project is conducting this onto-terminological work as an anchor point for a wider epistemological reflection to come, on the question of IT, placing at the heart of the question of this heritage, the history of a societal mutation and the emergence of new knowledge in terms of representations of the world, through the definition of this heritage and the new places of knowledge and practices. For this purpose, our multi-disciplinary team is committed to harnessing IT for its own benefit and, in its own image, combining approaches from the social sciences and humanities and the computer sciences. The new heritages and corpuses formed by data are leading us towards new approaches to terminology work. In doing so, we hope to provide an epistemological reflection not only on the subject of our study but also on the tools and methods related to language (linguistics and computer science).

Acknowledgments

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003). We thank also The QA Company, which is providing us with technical assistance as well as access to QAnswer https://qanswer.ai

References

[1] I. Kant, Critique de la raison pure, Aubier, Paris, 1997.
[2] C. Jacob, Rassembler la mémoire, volume 4, Diogène, 2001.
[3] Aristotle, L’Empiria, Librairie Philosophique J. Vrin, Paris, France, 2020.
[4] Gaudin, Socioterminologie: des problèmes semantiques aux pratiques institutionnelles, Presses universitaires de Rouen et du Havre, 1993.
[5] J. C. Boulanger, Présentation: images et parcours de la socioterminologie, Meta, 40.2, (1995) 194-205 (1995).
[6] C. Roche, M. Papadopoulou, Mind the Gap: Ontology Authoring for Humanists, in: Proceedings of JOWO: The Joint Ontology Workshops, Graz, Autriche, 2019.
[7] B. Russell, The philosophy of logical atomism, The Monist 28-29 (1918-1919) 495–527 and 32–63. Reprinted in Russell B., Logic and Knowledge, London, Allen and Unwin, 1956.
[8] H. Diels, Die fragmente der Vorsokratiker griechisch und deutsch, Weidmann, Berlin, 1903.
[9] C. Roche, Le terme et le concept : fondements d’une ontoterminologie, 2007. Actes TOTh 2007.
[10] F. Gaudin, Dire les sciences et décrire les sens: entre vulgarisation et lexicographie, le cas des dictionnaires de sciences, TTR: traduction, terminologie, rédaction 8 (1996) 11–27.

[11] H. Putnam, Raison, vérité et histoire, Editions de Minuit, Paris, 1984.
[12] C. Roche, Terminologie et ontologie, Langages (2005) 48–62.
[13] S. Auroux, Avant-propos, PUF, Paris, 1990, pp. vii–xx.
[14] A. Assal, et al., Sémantique et terminologie: sens et contextes, Terminologie et traduction (1992) 411–421.
[15] V. Claveau, M. C. L'Homme, Apprentissage par analogie pour la structuration de terminologie-utilisation comparée de ressources endogènes et exogènes, in: Actes de la conférence terminologie et intelligence artificielle (TIA-2005), 2005.
[16] J. Humbley, La traduction des métaphores dans les langues de spécialité: le cas des virus informatiques, Linx. Revue des linguistes de l'université Paris X Nanterre (2005) 49–62.
[17] M. C. L'Homme, Ressources lexicales, terminologiques et ontologiques: une analyse comparative dans le domaine de l'informatique, Revue française de linguistique appliquée (2008) 97–118.
[18] M. C. L'Homme, Terminology and lexical semantics, in: P. Faber, M. C. L'Homme (Eds.), Theoretical Perspectives on Terminology. Explaining terms, concepts and specialised languages, John Benjamins, Amsterdam/Philadelphia, 2022, pp. 237–259.
[19] M. C. L'Homme, Sélection de termes dans un dictionnaire d'informatique: Comparaison de corpus et critères lexicosémantiques, in: Actes Euralex 2004, 2004, pp. 583–593.
[20] M. C. L'Homme, Conception d'un dictionnaire fondamental de l'informatique et de l'internet: sélection des entrées, Le langage et l'homme 40 (2005) 137–154.
[21] V. Malaisé, Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels, Ph.D. thesis, Université Paris-Diderot - Paris VII, 2005.
[22] I. Meyer, Concept management for terminology. a knowledge engineering approach, in: P. Faber, M. C. L'Homme (Eds.), Theoretical Perspectives on Terminology. Explaining terms, concepts and specialised languages, John Benjamins, Amsterdam/Philadelphia, 2022, pp. 110–126.
[23] M. Rossi, Termes et métaphores, entre diffusion et orientation des savoirs, La linguistique 57 (2021).
[24] C. Djambian, M. Rossi, G. d'Ippolito, La médiation des objets aux savoirs scientifiques et techniques, in: Actes de la conference TOTh, Chambéry, 2023.
[25] N. González Granado, P. Drouin, A. Picton, De l'analyse statistique à l'apprentissage automatique : le langage R au service de la terminologie, Éla. Études de linguistique appliquée 208 (2022) 447–467.
[26] A. Kilgarriff, et al., The Sketch Engine : Ten years on, Lexicography 1 (2014) 7–36.
[27] P. Drouin, Term extraction using non-technical corpora as a point of leverage, Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 9 (2003) 99–115.
[28] S. Bird, E. Klein, E. Loper, Natural language processing with Python: Analyzing text with the Natural Language Toolkit, O'Reilly Media, Sebastopol, 2009.
[29] L. Anthony, Programming for corpus linguistics, in: M. Paquot, S. T. Gries (Eds.), A practical handbook of corpus linguistics, Springer, Cham, 2021, pp. 181–207.
[30] L. Lefeuvre, A. Condamines, MAR-REL : une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances (MAR-REL : a conceptual relation

markers database for Knowledge-Rich Contexts extraction), in: Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts, ATALA, Orléans, France, 2017, pp. 183–191.
[31] C. Djambian, Valorisation d'un patrimoine documentaire industriel et évolution vers un système de gestion des connaissances orienté métiers, Ph.D. thesis, Université Jean Moulin-Lyon III, 2010.
[32] F. Pan, J. R. Hobbs, Temporal aggregates in owl-time, in: FLAIRS, 2005, pp. 560–565.
[33] A. A. Salatino, et al., The computer science ontology: A comprehensive automatically-generated taxonomy of research areas, Data Intelligence 2 (2020) 379–416.
[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013). Available at https://arxiv.org/abs/1301.3781.
[35] P. Szekely, et al., Connecting the Smithsonian American Art Museum to the Linked Data Cloud, in: The Semantic Web: Semantics and Big Data, ESWC 2013, Berlin, Germany, 2013.
[36] V. de Boer, et al., Amsterdam museum linked open data, Semantic Web 4 (2013) 237–243.
[37] M. Doerr, et al., The Europeana Data Model (EDM), in: Actes de IFLA 76, Gottenburgh, Suède, 2010.
[38] R. Di Cosmo, Software heritage: why and how we collect, preserve and share all the software source code, in: 2018 IEEE/ACM 40th International Conference on Software Engineering, 2018.
[39] F. Clavaud, T. Francart, Sparnatural, un éditeur graphique souple et intuitif pour explorer des graphes de connaissances, in: Colloque Humanistica 2022, 2022.
[40] D. Diefenbach, et al., Towards a question answering system over the semantic web, Semantic Web 11 (2020).
[41] L. Smith, Heritage and its Intangibility, in: Ahmed Skounti y Ouidad Tebbaa: De l'immatérialité du patrimoine culturel, Bureau régional de l 'Unesco de Rabat, Marrakech, 2011, pp. 10–20.