



Universidade do Minho
Escola de Engenharia

João Rodrigues, pg46534
, Rolando Silva, pg46543
, Tomás de Sá, pg46544

**Extração (semi)Automática
de uma Ontologia com base
em Relatórios Arqueológicos**

Conteúdo

1	Introdução	1
2	Trabalho Relacionado	3
3	Conceção e Implementação do Sistema	5
3.1	Versão Automatizada	5
3.2	Versão inicial com a utilização de rules	6
3.3	Versão Final	8
3.4	Construção da Ontologia	10
4	Conclusões e trabalho futuro	12
4.1	Conclusões	12
4.2	Perspetiva de trabalho futuro	12

Lista de Figuras

1	Resultados da Versão Automatizada	6
2	Código do padrão	6
3	Output com a utilização de padrões	7
4	Código do Matcher full text	7
5	Output obtido pela utilização do Matcher	7
6	Output obtido da Versão Final 1	8
7	Output obtido da Versão Final 2	9
8	Output obtido da Versão Final 3	9
9	Localização da variável debug	10
10	Grafo representando a ontologia obtida	11
11	Localização da linha vis.show-buttons()	11

Capítulo 1

Introdução

Desde a antiguidade, o ser humano tenta demonstrar, representar e classificar, de uma forma abstrata, a si e a realidade que o rodeia. A esse estudo filosófico é denominado por ontologia.

Ontologia, no teor filosófico, é a ciência "do que é", dos tipos e estruturas de objetos, propriedades, eventos, processos e relações em todas as áreas da realidade. Remontando á Antiguidade Grega, ontologia era usada como sinónimo de metafísica, estudo do "que vem depois da Física"mas só em 1613 é que a palavra foi cunhada por dois filósofos, Rudolf Göckel e Jacob Lorhard.

Mas devido á evolução da tecnologia, as palavras vão ganhando novos significados e usos. Com a criação dos computadores e desenvolvimento das IAs, ontologia é "uma especificação de uma conceptualização", de acordo com Tom Gruber, um engenheiro americano conhecido por seu papel fundamental na definição de ontologias no contexto da inteligência artificial.

Tendo origem no início do séc.XXI, o termo "aprendizagem de ontologia"foi introduzido na campo da Inteligência Artificial como uma área de pesquisa que tem o objetivo de gerar automaticamente ontologias. Com uso de diferente técnicas e métodos podemos obter criar uma ontologia de forma semi-automática pois, geralmente, é necessário intervenção humana para uma categorização dos tipos de dados. Fornecendo ao modelo desenvolvido excertos de texto, dicionários, dados retirados de sites Web podemos obter outputs interessantes, mas com a possibilidade de refinar esse output utilizando ontologias pré-existentes e técnicas como estatísticas, utilização de regras (rules) e técnicas linguísticas, lógicas ou híbridas.Geralmente, uma ontologia consegue representar os dados como indivíduos, classes,atributos e seus relacionamentos.

Utilizando estas técnicas, o objetivo deste projeto será extrair (semi)automaticamente uma Ontologia com base em Relatórios Arqueológicos. Em relação ao projeto, a principal motivação do seu desenvolvimento foi o imenso material arqueológico recebido pela Universidade do Minho e este projeto tem como objetivo de ajudar na catalogação de documentos antigos, recentes e futuros de modo a preservar a informação e facilitar a sua pesquisa.

Neste documento serão apresentados as várias etapas do desenvolvimento do projeto, no Capítulo 2 será apresentado as pesquisas efetuadas e o seu aproveitamento teórico na conceptualização do modelo e no Capítulo 3, a Conceção e Implementação do sistema, onde serão descritas as técnicas e tecnologias utilizadas de modo a obter a ontologia desenvolvida ao longo do projeto.

Capítulo 2

Trabalho Relacionado

Vamos abordar agora o Trabalho Relacionado, devido a não sermos familiarizados no assunto de arqueologia, antes de se passar para o desenvolvimento da ontologia em si, foi necessário algum tempo para ser possível realizar alguma pesquisa e investigação sobre o assunto, que graças á ajuda do nosso professor orientador Orlando conseguimos identificar quais são as coisas importantes a se focar no tema.

Para iniciar começamos por analisar algumas dissertações de colegas, que mesmo não sendo sobre arqueologia, também realizavam a criação de ontologias sobre outros temas, o que nos permitiu ter uma boa base do que esperar do nosso trabalho, além de analisar como os mesmos realizavam todo o processo até á criação da ontologia.

Esta primeira dissertação era do colega, João Filipe da Costa Nunes, sobre a “Elaboração Semiautomática de uma Ontologia para Remédios a Partir de Textos Antigos de Medicina e de Culinária”. Durante a análise do trabalho, foi possível perceber melhor o que era as próprias ontologias, quais as suas estruturas, as suas várias metodologias, como avaliar a nossa ontologia e as suas diferentes técnicas de avaliação e quais as ferramentas e bibliotecas que poderíamos vir a usar e serem úteis no futuro, para o desenvolvimento da nossa própria ontologia. No final da análise foi possível ver, que existia um grande trabalho e complexidade para a criação de um ontologia, pois existe um grande pré-processamento, tratamento de texto e avaliação.

A nível de investigação na internet, investigamos e lemos alguns artigos para que fosse possível aumentar o conhecimento e o vocabulário na área, para que seja possível compreender os termos científicos¹, pois era uma nova temática que nenhum de nós estava familiarizado com ela. Também investigamos alguns artigos relativos á construção de uma ontologia ² e ainda sobre a criação de ontologias utilizando a linguagem Python, que seria essa a linguagem que seria usada ³. Para além disso, também tivemos acesso a alguns exemplos de textos (“sondagens”), que tinham sido feitas por arqueólogos, que o nosso professor orientador Orlando, nos conseguiu enviar, onde estas sondagens viriam depois a ser usados como testes para a nossa ontologia que estávamos a criar, que já com todo o trabalho de inves-

tigação e pesquisa realizado foi possível já compreender alguns termos, conceitos, relações usadas e a forma de como a divisão do texto para cada escavação era feita.

A partir de todo o conhecimento que tivemos depois de todo a análise e leitura destes documentos, tentamos criar uma ontologia, onde foi aí que nos deparamos com alguns problemas, já tínhamos algum processamento de texto, a nível de verbos, sinónimos, entre outros, mas não tínhamos o conhecimento para saber quais os atributos e relações que eram importantes nos focar no processamento do texto.

Com o decorrer das reuniões o professor orientador Orlando, achou que o tema da tese de um outro colega fazia mais sentido com nosso, que era a tese do José Pedro Saraiva de Carvalho, “Uma Ontologia para o Livro das Propriedades da Mesa Arcebispal de Braga”, onde a forma de como era feita o processamento do texto era parecido á forma de como fazíamos na arqueologia. Também foi realizada uma reunião com um grupo de arqueólogas da Universidade do Minho para compreendermos melhor os termos, atributos e as relações importantes num texto de arqueologia de uma escavação que eram importantes e deveríamos nos focar. Com esta reunião foi possível saber da existência de um documento “Definition of the CIDOC Conceptual Reference Model”, que tinha a informação sobre o estabelecimento de relações espaciais, envolvendo questões como locais, atores, objetos, entre outros, que são elementos importantes na nossa ontologia, pois podemos nos basear nisto para que seja possível realizar a extração, destes conceitos e as suas relações a partir de um conjunto de textos. Também foram fornecidos alguns outros relatórios de escavação e outros documentos⁴.

Com o decorrer deste tempo foi possível melhorar a nossa ontologia original, onde já tínhamos iniciado a implementação de rules que nos permitia adicionar padrões que nos permitia reconhecer as camadas de terreno, numa fase inicial como era de suspeitar os resultados não foram muitos otimizados, mas já eram bastante satisfatórios. Agora a partir da descrição das entidades do CIDOC, o objetivo era criar uma ou mais rules para conseguir obter a informação a partir do texto que for dado, como por exemplo a entidade E52 Time-Span ver a descrição e os exemplos e fazer uma rule para isso.

Mas todo o processo e passos desde a escolha do modelo, pré-processamento até á criação da ontologia final será explicado no capítulo seguinte de uma forma mais detalhada.

Capítulo 3

Conceção e Implementação do Sistema

Neste próximo tópico, serão descritas as etapas do trabalho prático desenvolvidas ao longo do projeto. Na realização deste trabalho foram utilizados modelos e como falado anteriormente o trabalho foi evoluindo, com a adição de novas técnicas e recursos. Com a progressão alcançada no projeto, pode-se dividir esta evolução em 4 etapas, na qual foram nomeadas como:

- Automatizada;
- Recorrendo à utilização de rules;
- Versão Final do processamento (versão melhorada da segunda etapa);
- Construção da Ontologia (grafo).

3.1 Versão Automatizada

A versão automatizada, já foi um pouco abordada anteriormente, e consistia numa versão muito simples de análise e processamento de texto. Este protótipo tinha como função de quando lhe era fornecido um texto, este iria analisar cada palavra do mesmo e catalogar as palavras como estas sendo verbos ou nomes. Como podemos ver é uma técnica bem simples de análise de textos, mas que já nos permite abordar a temática de processamento, que é uma temática bastante importante onde, nas próximas fases será bem explorada, para que seja possível criar as relações. Tal como podemos ver na imagem seguinte, os resultados ainda não são aqueles desejados.


```

Noun phrases: ['A sondagem', 'uma forma retangular', 'cujas dimensões', 'm.', 'Os trabalhos', 'a decapagem', 'camada superficial', '(UE016', 'que', 'elementos laterícios', 'nódulos', 'argamassa', 'vesti-  
gios', 'aterros', 'obras', 'Uma vez', 'a camada', '(UE016', 'dois enchimentos', '(UE026', 'que', 'duas estruturas', 'que', 'um muro', 'granito', 'orientação', 'O-E', 'uma canalização', 'recurso', 'mem  
o material', 'que', 'anterior (UE020', 'fim', 'outro enchimento', '(UE025', 'a arena granítica', '(UE024', 'A intervenção', 'se', 'o nível geológico', 'as estruturas', 'situ', 'uma altitude média', 'á  
rea saibrosa', 'ruínas', 'O espólio', 'sondagem', 'apenas na camada superficial', '(UE016', 'uma variedade significativa', 'produções', 'facto', 'os fragmentos cerâmicos', 'fabricos comuns', 'período b  
aixo', 'assim como material', 'construção', 'época recente', 'e ainda um objeto', 'metal', 'cuja função', 'forma', 'A sondagem', 'uma sequência estratigráfica', 'contextos', 'anteriores', 'feito',  
os enchimentos recentes', '(UE026', 'duas estruturas', 'que', 'a ocupação', 'área', 'periferia', 'cidade', 'Braga', 'um muro', 'uma canalização', '(UE020', 'Mau', 'a ausência', 'materiais provenientes  
da generalidade das camadas', 'essas estruturas', 'transformação', 'áreas extramuros', 'Braga', 'grandes quintas', 'que', 'a paisagem', 'a exploração agrícola', 'terras férteis', 'vales', 'Este', 'pri  
ncipalmente, do Cávado', 'em época moderna', 'período', 'uma reestruturação', 'sistema', 'abastecimento', 'água', 'com a ampliação', 'rede', 'conduções', 'da implantação', 'poços', 'a difusão', 'plan  
tio', 'milho', 'mido']

Verbs: ['implantar', 'apresentar', 'iniciar', 'apresentar', 'associar', 'remover', 'individualizar', 'recobrir', 'preservar', 'poder', 'identificar', 'construir', 'desenvolver', 'individualizar', 'depos  
itar', 'dar', 'finalizar', 'identificar', 'preservar', 'identificar', 'encontrar', 'revelar', 'associar', 'puder', 'determinar', 'apresentar', 'identificar', 'preservar', 'assinalar', 'tratar se', 'grar  
, 'acreditar', 'associar', 'moldar', 'assister se', 'relacionar']

UE026 LOC  
O-E MISC  
Braga LOC  
Este MISC  
Cávado LOC

Noun phrases: ['A sondagem', 'dimensões', '6x14', 'm.', 'uma forma retangular', 'A intervenção', 'remoção', 'enchimentos', 'que', 'as UE022', 'que', 'inclusões', 'material', 'construção', 'cerâmico',  
'vez', 'dois muros', 'alvenaria irregular', 'granito', 'que', 'oeste', '(UE028', 'a este', '(UE030', 'um pavimento', 'lajes', 'granito', '(UE021', 'marcas', '(UE027', 'trecho', 'cerca de 13m.', 'Os trab  
alhos', 'a preservação', 'referida calçada', 'cuja altitude', 'os 156,05m', 'O espólio', 'sondagem', 'apesar de restrito', 'camadas', 'UE022', 'nível material', 'com a presença', 'cerâmicas', 'vi  
dros', 'no', 'que', 'produções', 'feito', 'enchimento UE022', 'objetos olizros', 'fabrico comum', 'época moderna', 'nível', 'elementos metálicos', 'função indeterminada', 'uma moeda', 'Por', 'vez', 'o e  
spólio proveniente', 'camada UE023', 'fragmentos cerâmicos', 'produção comum', 'cronologia moderna', 'construção', 'assim como vidros incolores', 'Os trabalhos', 'sondagem', 'o registo', 'uma sequênci  
a estratigráfica', 'uma calçada', 'lajeados', 'granito', '(UE021', 'dois muros', 'mesmo material', '(UE028', 'dois robustos enchimentos', '(UE022', 'que', 'essa via', 'Mau', 'os níveis', 'implantação',  
'calçada', 'que', 'situ', 'as camadas', 'que', 'o abandono', 'estrutura', 'uma utilização', 'um período bastante tardio', 'a presença', 'materiais', 'época contemporânea', 'elemento', 'que', 'o seu f  
uncionamento', 'longo', 'período', 'exploração', 'antiga quinta', 'Portas', 'Cangosta d'Abraão', 'Planta Topográfica de Francisque Goullard, de 1883/84']

Verbs: ['implantar', 'apresentar', 'começar', 'sobrepunhar', 'individualizar', 'apresentar', 'decapar', 'pôs-se', 'descobrir', 'delimitar', 'constituir', 'rodar', 'preservar', 'dar', 'terminar', 'referi  
r', 'variar', 'recolher', 'diversificar', 'tocar', 'identificar', 'exumar', 'cunhar', 'apresentar', 'realizar', 'permitir', 'representar', 'delimitar', 'constituir', 'selar', 'grar', 'intervencionar',  
preservar', 'salientar', 'documentar', 'aludir', 'dar', 'sugerir', 'associar', 'referenciar']

UE022 MISC  
UE022 MISC  
UE022 MISC  
UE023 LOC  
UE028 MISC  
UE022 MISC  
Cangosta MISC  
Planta Topográfica de Francisque LOC

```

Figura 1: Resultados da Versão Automatizada

3.2 Versão inicial com a utilização de rules

A próxima fase, já é algo um pouco mais complexo, pois vamos tentar identificar as palavras e dizer qual o papel dela na área de arqueologia, por exemplo, se aparecer uma palavra "vaso", o programa ao processar essa mesma palavra saber que se trata de um objeto físico. Para isto já foi necessário a utilização de rules para se implementar os padrões, para que o nosso programa consiga realizar o processamento, tal como descrito no exemplo. Temos aqui, o exemplo do código e dos resultados do mesmo numa primeira implementação de padrões, onde esta tinha objetivo realizar o reconhecimento do tipo de terrenos.

```

102
103 ## Alterações do modelo
104 ruler = nlp.add_pipe("entity_ruler")
105
106 pattern_camada = [{'LEMMA': 'camada'},
107                    {'POS': 'ADJ'}]
108
109 patterns = [{"label": "CAMADA", "pattern": pattern_camada}]
110
111 ruler.add_patterns(patterns)
112
113

```

Figura 2: Código do padrão

```

camada humosa CAMADA
UEs002 ORG
UEs003 ORG
UE009 LOC
camada areno-limo-argilosa CAMADA
UEs004 ORG
camada abundante CAMADA
UE003 MISC
UE004 MISC
UEs004 MISC
camada geológica CAMADA
camada superficial CAMADA

```

Figura 3: Output com a utilização de padrões

Também foi tentado realizar esta implementação, da mesma forma, mas por outros métodos, como utilizando o “Matcher full text”.

```

print("Matcher full text")
expression = "camada ([^ ]*)"
for match in re.finditer(expression, doc.text):
    start, end = match.span()
    span = doc.char_span(start, end)
    # This is a Span object or None if match doesn't map to valid token sequence
    if span is not None:
        print(span.text, "CAMADA")

```

Figura 4: Código do Matcher full text

```

Matcher full text
camada humosa CAMADA
camada areno-limo-argilosa CAMADA
camada abundante CAMADA
camada geológica CAMADA
camada (UE005) CAMADA
camada superficial CAMADA

```

Figura 5: Output obtido pela utilização do Matcher

No final, analisando os resultados obtidos nestas experiências já é alguma coisa, pois alguns dos resultados que são apresentados são bastantes satisfatórios, contudo outros não são, daí mostrar que é necessário ainda um polimento. Mas permitiu abrir as portas na implementação de rules e padrões, que depois serão usados não só para identificação do tipo de terreno, como também para identificar os atores presentes, objetos encontrados, locais, entre outros.

3.3 Versão Final

De seguida a nossa próxima fase é fazer o que foi feito anteriormente, só que para outros atributos relevantes que depois, serão usados na construção da ontologia tendo em conta as suas relações. Sendo estes os Atores, os vários movimentos, os objetos físicos encontrados, coisas físicas, o período e ainda o local. Tal como podemos ver na imagem seguinte. (As imagens seguintes são apenas excertos dos resultados e não demonstram todo o resultado, são apenas para que seja possível mostrar exemplos de cada tópico).

```
Ontology:

E39 Actor
  Ã
  UEs004
  Ã
  Mau
  Ã s camadas
  Ã s produÃ§Ães

E9 Move
  implantar
  arraigar
  arvorar
  enraizar
  enxerir
  estabelecer
  firmar
  fixar
  hastear
  içar
  inserir
  plantar
  iniciar
  abrir
  catequizar
```

Figura 6: Output obtido da Versão Final 1

```

E19 Physical Object
    moeda

E18 Physical Thing
    camada humosa
    camada areno-limo-argilosa
    camada abundante
    camada abastado
    camada copioso
    camada diluviano
    camada fecundo
    camada fértil
    camada numeroso
    camada opulento
    camada produtivo
    camada profuso

```

Figura 7: Output obtido da Versão Final 2

```

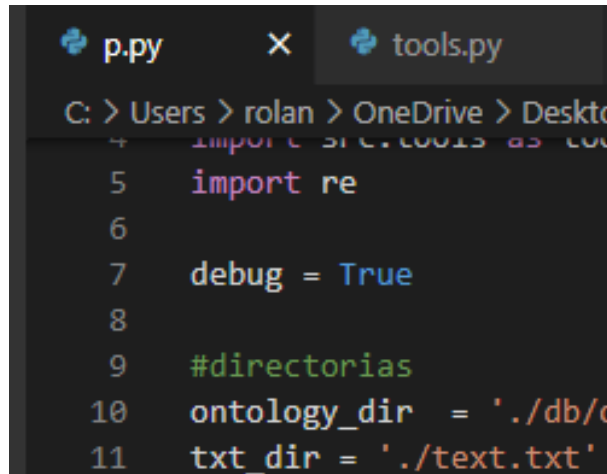
E4 Period

E53 Place
    UE009
    cerâmica
    sequência
    rio Este
    UE005
    UE004
    UE004).Por
    UEs026
    UE019
    UE025
    espalio
    UEs016
    Braga
    Cívado

```

Figura 8: Output obtido da Versão Final 3

Esta informação aparece no terminal, contudo por não se achar que se encontra de uma forma organizada, foi implementado uma pequena funcionalidade, onde toda esta informação e ainda mais (o texto que foi processado, as frases nominais, todos os tokens e todas as entidades), é disponibilizada num ficheiro de texto ("debug.txt"), para isso basta no código basta escrever "debug = True" em vez de "debug = False".



```
p.py X tools.py
C: > Users > rolan > OneDrive > Desktop
4 import re
5 import re
6
7 debug = True
8
9 #directorias
10 ontology_dir = './db/o
11 txt_dir = './text.txt'
```

Figura 9: Localização da variável debug

3.4 Construção da Ontologia

Por fim, com todas este processamento de texto é possível criar um grafo que representa a ontologia, tendo em conta as relações. O script cria um ficheiro ".html" que tem o nome de "ontology.html", onde está representado o grafo.

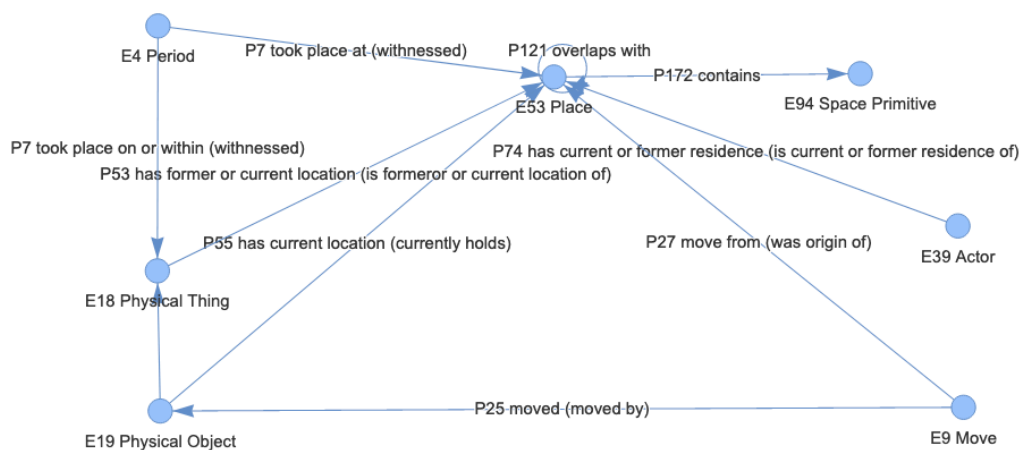


Figura 10: Grafo representando a ontologia obtida

Quando aberto o ficheiro, permite que o utilizador interage com ele, conseguindo mover e ampliar á sua maneira, para tal basta remover de comentário a linha onde tem "vis.show-buttons()" no ficheiro "tools.py" e correr o código.

```

126
127     # Adicionar o grafo
128     vis.from_nx(G)
129     vis.barnes_hut()
130
131     #vis.show_buttons()
132
133     # Mostra o grafo
134     vis.show("ontology.html")

```

Figura 11: Localização da linha vis.show-buttons()

Capítulo 4

Conclusões e trabalho futuro

4.1 Conclusões

Graças as ontologias, conseguimos relacionar vários conceitos e obter essas relações, tornando assim mais fácil a sua pesquisa e o acesso à informação. Como pode-se notar, pode-se aplicar em qualquer área que exige grande processamento de texto pois com a utilização de várias técnicas obtém-se uma grande capacidade de adaptação do modelo. Na área da arqueologia, com o passar dos tempos, há imensa informação que entra no esquecimento, podendo levar com ela fatos importantes sobre a nossa história, métodos e instrumentos utilizados nas escavações, os autores da escavação, períodos de tempo onde certos objetos se localizam. Mas com este trabalho, poderemos facilitar o acesso à informação, desenterrando arquivos mais antigos e obter uma maior eficiência na sua obtenção e pesquisa.

Este projeto passou por várias fases até a concepção da versão final obtida. Com o auxílio de trabalhos anteriores e pesquisas na Web, conseguimos ter um ponto de partida teórico e prático no processo de elaboração da ontologia e com algumas reuniões com profissionais da área da arqueologia, foram revelados mais alguns atributos relevantes que o modelo deveria retirar dos relatórios arqueológicos.

O projeto sofreu grandes evoluções ao longo do desenvolvimento, desde de uma versão totalmente automatizada, onde se obteve um output significativo para o início mas muito simples até a utilização de técnicas como a adição de padrões e regras para conseguir detectar, com maior precisão, atributos mais relevante na área arqueológica como períodos de tempo, autores e locais.

4.2 Perspetiva de trabalho futuro

Com o trabalho efetuado neste projeto, a ontologia adquirida deveria evoluir mais. Com mais pesquisa sobre a área, podíamos aplicar mais técnicas que nos permitissem alcançar uma maior precisão e refinamento dos atributos. O projeto produzido está bem encaminhado sendo necessário mais algum trabalho

mas seria essencial a criação de uma interface entre o utilizador e a ontologia que pudesse facilitar as suas interações pois nem todos os utilizadores poderão usufruir das suas funcionalidade no estado atual. Essa interface poderia efetuar pesquisas de conceitos e atributos mais específicos ou apresentar as várias relações obtidas no processamento de um texto. Assim, com uma interface mais user-friendly, provavelmente disponível na Web em forma de um site ou um API, mais pessoas poderiam utilizar, no que resulta no enriquecimento do conhecimento no futuro, em melhores trabalhos produzidos focados na área e melhor acesso á informação que possivelmente já não era utilizada.

Bibliografia

- A domain-specific formal ontology for archaeological knowledge sharing and reusing. URL <https://www.semanticscholar.org/paper/A-Domain-Specific-Formal-Ontology-forKnowledge-and-Zhang-Cao/fa33b999fae6f80aa15557063e2840f2cd8a220f>.
- Ontology development. URL https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.
- Building ontologies with python. URL <https://paul-bruffett.medium.com/building-ontologies-with-python-84238d6eee5>.
- Relatório arqueológicos. URL <http://orcp.hustoj.com/wp-content/uploads/2015/10/2009-Ontologies-for-Cultural-Heritage.pdf>https://link.springer.com/chapter/10.1007/3-54036277-0_20.