

Gerenciamento da Privacidade baseado em Ontologia em Sistemas de rastreamento de Saúde usando Privacidade Diferencial

Erika Guetti Suca

Universidade de São Paulo

16 de junho de 2023

Agenda

- 1 Introdução
- 2 Conceitos Fundamentais
- 3 Trabalhos Relacionados
- 4 Proposta
- 5 Bibliografia

Introdução

Problema

Apreender informação dada uma base de conhecimento que representa determinada informação de um grupo de participantes sem associar dados sensíveis de um membro à sua identidade.

- O problema é evitar o vazamento de dados sensíveis coletados para análise.
- Dados que isoladamente poderia ser inofensivos. Porém, quando combinados podem comprometer a privacidade dos participantes.

Conceitos Fundamentais

Ontologias

Introdução

Conceitos
Fundamentais

Trabalhos
Relacionados

Proposta

Bibliografia

- Uma ontologia é uma **teoria lógica** para **estruturar** o significado pretendido de um vocabulário formal, i.e., um compromisso ontológico com uma conceitualização particular do mundo [Gua98].
- Conjunto de axiomas, relações de subsunção e subordinação entre classes e propriedades. Os axiomas fazem possíveis as afirmações e as subsunções fazem possíveis que se estabeleçam as equivalências e as classes. São teorias lógicas **consistentes e coerentes** [SS04].
- São instrumentos para **representação do conhecimento** atuando, principalmente, no controle terminológico.

Conceitos Fundamentais

Porque usamos ontologias?

- Permitem a **estruturação de dados abertos conectados** de alta qualidade. Dados livres e compartilhados na Internet, para uso de qualquer pessoa ou máquina, permitindo o cruzamento de diferentes fontes, para serem livremente reutilizados pela sociedade [IB15].
- Conectam silos de dados, pessoas, lugares e coisas. São particularmente úteis quando há incerteza nos dados ou há relações complexas entre os processos [KM19].
- Ontologias têm auxiliado na modelagem da semântica de conceitos médicos e facilitado a troca de dados médicos entre diversas terminologias.

Conceitos Fundamentais

Representação de Ontologias

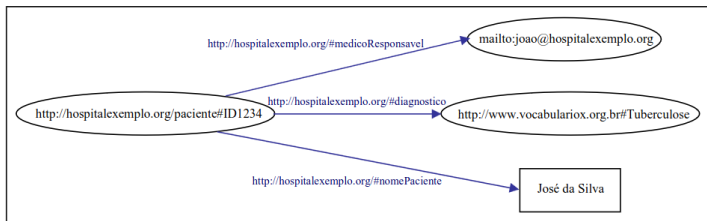


Figura: Exemplo de tripla RDF (Resource Description Framework) [dCM09].

Conceitos Fundamentais

Introdução

Conceitos Fundamentais

Trabalhos Relacionados

Proposta

Bibliografia

Art 5. Marco Civil da Internet Brasil - Lei Geral de Proteção de Dados Pessoais

Dado pessoal: dado relacionado à pessoa natural identificada ou identificável, inclusive através de números identificativos, dados locais ou identificadores eletrônicos.

Dados sensíveis: dados pessoais que revelam a origem racial ou étnica, as convicções religiosas, filosóficas ou morais, as opiniões políticas, a filiação a sindicatos ou organizações de caráter religioso, filosófico ou político, dados referentes à saúde ou à vida sexual, bem como dados genéticos.

Dado anonimizado: dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento.

Conceitos Fundamentais

Art 5. Marco Civil da Internet Brasil - Lei Geral de Proteção de Dados Pessoais

Tratamento: conjunto de ações referentes a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, transporte, processamento, arquivamento, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração [dJ19].

Conceitos Fundamentais

Art 5. Marco Civil da Internet Brasil - Lei Geral de Proteção de Dados Pessoais

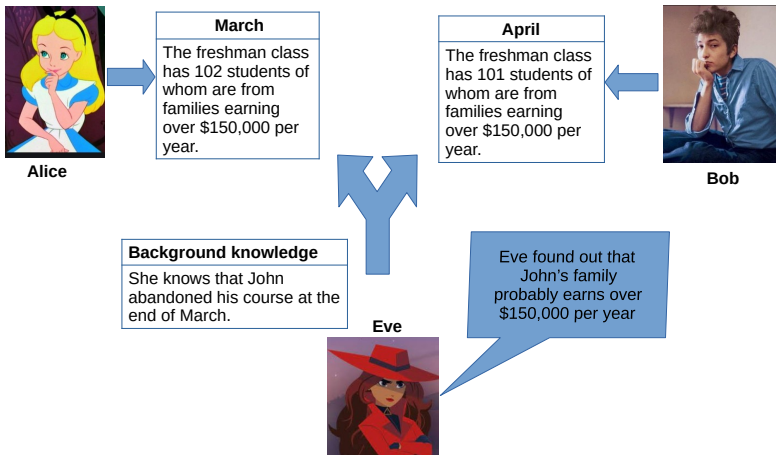
Privacidade de dados

A privacidade de dados é um estado de proteção de dados focado no tratamento, compartilhamento e uso adequado de dados pessoais, sensíveis e anonimizados para gerenciar riscos relacionados à exposição inadequada (Adaptação de [dJ19]).

Conceitos Fundamentais

O problema da composição

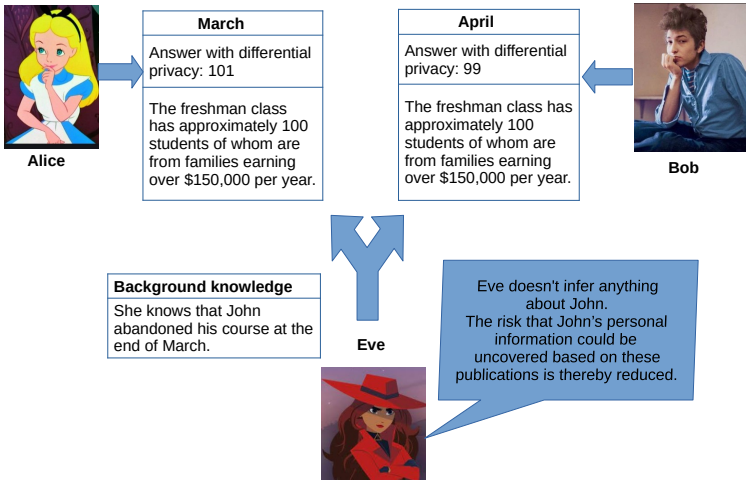
Exemplo adaptado de [WANV20].



Conceitos Fundamentais

O Problema da Composição

Exemplo com aplicação da privacidade diferencial.



Trabalhos Relacionados

Privacidade Diferencial

Com base em respostas aleatórias e ajustadas [EPK14].

Vantagens:

- Fortes garantias de privacidade para cada participante.

Desvantagens:

- Sua proposta não foi adaptada para ontologias OWL 2.

Trabalhos Relacionados

Anonimização

Baseado em técnicas que restringem a quantidade de dados liberados, usando funções que generalizam ou suprimem dados [CGH08], [GK16].

Vantagens:

- Visualizações admissíveis.

Desvantagens:

- Eles não modelam o conhecimento prévio do usuário sobre uma base de conhecimento para criar as visualizações seguras possíveis.

Trabalhos Relacionados

Conformidade com as Políticas de Segurança

Restringindo o acesso a dados confidenciais [EKP11], [GM14], [BSP14, BPS15], [BBN17, BKN19].

Vantagens:

- Construção de visualizações seguras.
- Conformidade e segurança com as políticas.

Desvantagens:

- Reparar ontologias para se livrar de consequências indesejadas ainda não é suficiente, pois pode ser que um possível invasor possua informações relevantes de outras fontes.

Conceitos Fundamentais

Privacidade Diferencial

A privacidade diferencial é um conjunto de algoritmos para compartilhar informações, preservando a privacidade dos indivíduos no conjunto de dados com a seguinte condição [Dwo06].

$$\forall t \in \text{Range}(\mathcal{A}) : \frac{\Pr(\mathcal{A}(\mathcal{D}) = t)}{\Pr(\mathcal{A}(\mathcal{D}') = t)} \leq e^\epsilon \quad (1)$$

O parâmetro ϵ controla quanta informação pode vaziar regulando quantos dados sintéticos aleatórios podem ser introduzidos.

Conceitos Fundamentais

Mecanismo de Laplace

O mecanismo Laplace é usado para consultas com a seguinte estrutura[DR14]:

- Quantos elementos no conjunto de dados satisfazem uma propriedade P ?
- Qual das consultas de contagem M tem o maior/menor/etc. valor?
- Qual das consultas de contagem M tem um maior/inferior/etc. valor do que o valor P ?

Conceitos Fundamentais

Introdução

Conceitos
Fundamentais

Trabalhos
Relacionados

Proposta

Bibliografia

O Mecanismo Exponencial

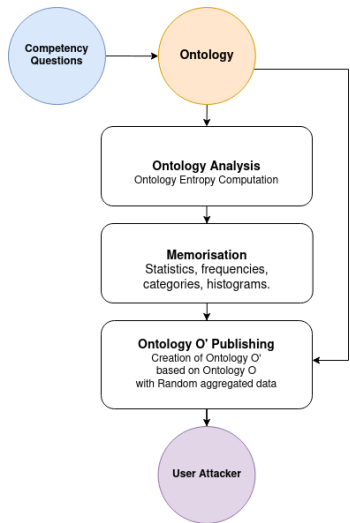
O mecanismo Exponencial é aplicado para consultas não numéricas que podem ser sensíveis a pequenas perturbações [DR14], como:

- Qual é a doença mais comum na cidade de São Paulo?
- Qual é o melhor preço em um leilão?
- Qual é o horário ideal para evitar o trânsito em São Paulo?
- Qual é a resposta ótima para o tempo de espera?

Propomos um modelo para adquirir informações a partir de uma ontologia controlando a quantidade de informações e limitando a capacidade de um invasor aprender dados confidenciais do usuário.

Metodologia

Mecanismos propostos para preservar a privacidade



Metodologia

Mecanismo: Respostas com dados sintéticos agregados

Um algoritmo \mathcal{A} satisfaz ϵ -privacidade diferencial, onde $\epsilon \geq 0$, se e somente se para qualquer ontologia \mathcal{O} e \mathcal{O}' que diferem pelo menos uma instância, temos [Dwo06]:

$$\forall \mathcal{T} \subseteq \text{Range}(\mathcal{A}) : \Pr[\mathcal{A}(\mathcal{O}) \in \mathcal{T}] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{O}') \in \mathcal{T}], \quad (2)$$

- onde $\text{Range}(\mathcal{A})$ denota o conjunto de todas as saídas possíveis do algoritmo \mathcal{A} .
- ϵ é o parâmetro denominado orçamento de privacidade. Geralmente é recomendado usar os valores de 0,01, 0,1, $\ln 2$ e $\ln 3$ [DS10].

Metodologia

Mecanismo de Laplace

Seja \mathcal{M}_L um algoritmo que gere uma resposta com dados aleatórios agregados dada uma consulta f e uma ontologia \mathcal{O} com sensibilidade Δf . Então,

$$\mathcal{M}_L(\mathcal{O}) = f(\mathcal{O}) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) \text{ satisfaz } \epsilon\text{-DP.} \quad (3)$$

$\text{Lap}\left(\frac{\Delta f}{\epsilon}\right)$ é ϵ -DP.

Mecanismo de Laplace

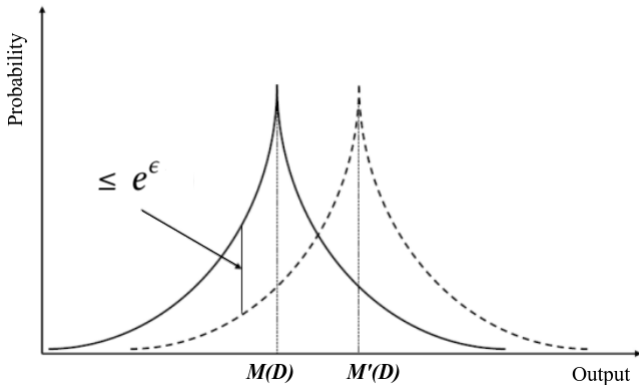


Figura: Privacidade diferencial via perturbação de Laplace.

Metodologia

Mecanismo Exponencial

Seja \mathcal{M}_E um algoritmo que produza para cada elemento $x \in \phi$ uma probabilidade proporcional a

$$\mathcal{M}_E(\mathcal{O}, \phi) \propto \exp\left(\frac{\epsilon u(\mathcal{O}, \phi)}{2\Delta u}\right) \quad (4)$$

- ϕ é a saída da consulta $f(\mathcal{O})$, ϵ é um parâmetro de privacidade, $u(\mathcal{O}, \phi)$ é uma função de pontuação, e, Δu é a sensibilidade da pontuação da função u .
- A intuição é normalizar os escores ajustando os valores medidos em diferentes escalas para uma escala teoricamente comum.

Exemplo

Monitoramento do Ciclo Menstrual

Revelar informações muito íntimas, como o histórico menstrual e/ou sexual, poderiam provocar sérios prejuízos para as usuárias.

Coleta de dados biofísicos por tecnologias de rastreamento de saúde.

- Tecnologias que enfatizam os desafios pessoais.
- Com base em auto-quantificação e técnicas de auto-educação.

Exemplo

Monitoramento do Ciclo Menstrual







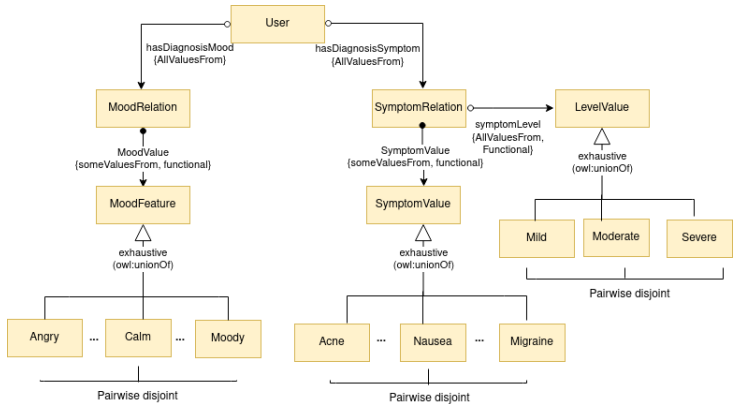
 Sleep duration	 Sex
	<input type="button" value="Had sex"/> <input type="button" value="Multiple sessions"/>
 Feeling stressed?	<input type="button" value="No sex"/>
<input type="button" value="Yes"/> <input type="button" value="No"/>	
 Spotting	 Cervical mucus (CM) 
	CM TEXTURE
<input type="button" value="Light"/> <input type="button" value="Medium"/>	<input type="button" value="Dry"/> <input type="button" value="Sticky"/>
<input type="button" value="Heavy"/> <input type="button" value="None"/>	<input type="button" value="Watery"/> <input type="button" value="Raw Eggwhite"/>
 Weight 54 KG	<input type="button" value="Creamy"/>
 Basal body temperature (BBT)  36 °C	CM AMOUNT
	<input type="button" value="Light"/> <input type="button" value="Medium"/> <input type="button" value="Heavy"/>

Figura: Dados coletados e analisados diariamente, mais de 3 milhões de usuárias [Owe16].

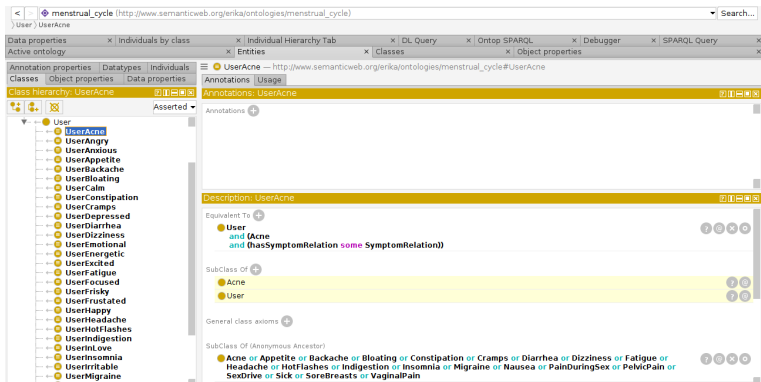
Exemplo

Ontologia proposta usando os padrões de ontologia *N-ary* e *Value Partition*.



Exemplo

Modelamento no Protegé



Exemplo

Monitoramento do Ciclo Menstrual

Questões de competência propostas para ontologia.

Consultas de contagem (Mecanismo de Laplace):

- Quantos elementos na Ontologia \mathcal{O} satisfazem a propriedade P ?
- Quantos usuários tiveram insônia em setembro?
- Quanto é o número médio de dias que as mulheres estão tentando engravidar com mais de 35 anos?
- Quantas mulheres estão tentando engravidar depois dos 40?

Exemplo

Monitoramento do Ciclo Menstrual

Questões de competência propostas para ontologia.

Consultas altamente sensíveis a pequenas quantidades
(Mecanismo Exponencial):

- Qual é o sintoma físico mais comum na cidade de São Paulo?
- Qual é a data de início do próximo período?
- Qual é o dia mais provável de engravidar?

Exemplo

Respondendo a perguntas com privacidade diferencial

Q1) Quantos usuários tiveram acne ou apetite ou insônia ou dor lombar ou inchaço ou Constipação ou enxaqueca ou câibras ou náusea ou ficou doente em setembro?

SPARQL Query:

```
3 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
5 PREFIX mens:<http://www.semanticweb.org/erika/ontologies/menstrual_cycle#>
6 SELECT (count(distinct ?users) as ?users_with_symptoms)
7 WHERE {
8
9     ?users rdfs:type ?user .
10    ?user owl:intersectionOf ?list .
11    ?list rdfs:rest*/rdfs:first ?equivalentClass .
12
13    ?equivalentClass owl:intersectionOf ?list2.
14
15    ?list2 rdfs:rest*/rdfs:first ?s.
16
17    FILTER ( ?s = mens:Acne || ?s = mens:Appetite || ?s = mens:Insomnia
18            || ?s = mens:Backache || ?s = mens:Bloating || ?s = mens:Constipation
19            || ?s = mens:Migraine || ?s = mens:Cramps || ?s = mens:Nausea || ?s = mens:Sick)
20
21 }
```

Exemplo

Respondendo a perguntas com privacidade diferencial

Q1) Quantos usuários tiveram acne ou apetite ou insônia ou dor lombar ou Inchaço ou Constipação ou Migraine ou Cãibras ou Náusea ou Doente em setembro?

Soma verdadeira, \mathcal{A} : 69

```
"69"^^<http://www.w3.org/2001/XMLSchema#integer>
```

Com privacidade diferencial, \mathcal{A}' : 71

$d \sim \text{Lap}(\frac{1}{\epsilon})$, $e^{d/\beta} = e^\epsilon$, $d = 1$, e $c' = c + d = 2$

Exemplo

Respondendo a perguntas com privacidade diferencial

Estatísticas memorizadas: funções de pontuação.

	Users	$F(U)$
User1	15-19	0.05
User2	20-24	0.10
User3	25-29	0.20
User4	30-34	0.30
User5	35-39	0.20
User6	40-44	0.18
User7	44-49	0.02

	Symptoms	$F(Symptom)$
Symptom1	Acne, Appetite	0.05
Symptom2	Backache, Bloating	0.10
Symptom3	Constipation	0.30
Symptom4	Cramps, Diarreheia	0.25
Symptom5	Nausea, Sick	0.30

Exemplo

Respondendo a perguntas com privacidade diferencial

Q2) Qual é o sintoma físico mais comum nesta cidade?

Options	Number of users	$u(\mathcal{O}, \phi)$	$\epsilon = \ln 2$	$\epsilon = 1.0$
Insomnia	1	0.025	0.15	0.09
Acne	14	0.350	0.35	0.36
Migraine	25	0.625	0.50	0.55

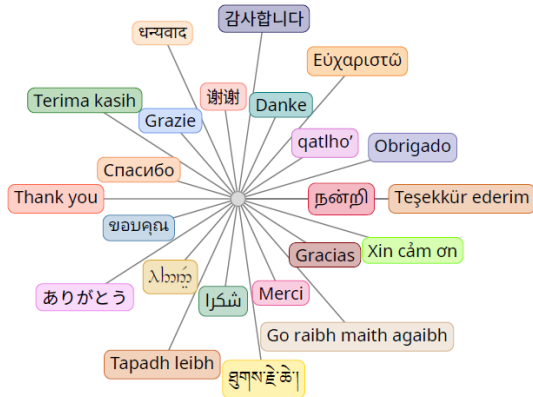
$$\mathcal{M}_E(\mathcal{O}, \phi) \propto \exp\left(\frac{\epsilon u(\mathcal{O}, \phi)}{2\Delta u}\right)$$

Adotamos o número de usuários em cada sintoma dividido pelo número total de usuários como a função de pontuação $u(\mathcal{O}, \phi)$. Como a exclusão de um usuário terá um impacto máximo de 1 no resultado de u , então a sensibilidade de u é $\Delta u = 1$.

Comentários finais

Podemos citar algumas vantagens do nosso modelo:

- Um atacante não tem nada a aprender sobre um determinado indivíduo.
- Retornar o valor verdadeiro não é problema, o objetivo é não associar informações confidenciais à identidade de um participante.
- Como próximo passo, iremos implementar e experimentar nossa proposta com grandes ontologias.



Bibliografia I



Franz Baader, Daniel Borchmann, and Adrian Nuradiansyah, *The identity problem in description logic ontologies and its application to view-based information hiding*, Semantic Technology - 7th Joint International Conference, JIST 2017, Gold Coast, QLD, Australia, November 10-12, 2017, Proceedings, 2017, pp. 102–117.



Franz Baader, Francesco Kriegel, and Adrian Nuradiansyah, *Privacy-preserving ontology publishing for EL instance stores*, Logics in Artificial Intelligence - 16th European Conference, JELIA 2019, Rende, Italy, May 7-11, 2019, Proceedings, 2019, pp. 323–338.

Bibliografía II



Piero A. Bonatti, Iliana M. Petrova, and Luigi Sauro, *Optimized construction of secure knowledge-base views.*, Description Logics (Diego Calvanese and Boris Konev, eds.), CEUR Workshop Proceedings, vol. 1350, CEUR-WS.org, 2015.



Piero A. Bonatti, Luigi Sauro, and Iliana M. Petrova, *A mechanism for ontology confidentiality*, Proceedings of the 29th Italian Conference on Computational Logic, Torino, Italy, June 16-18, 2014., 2014, pp. 147–161.



Bernardo Cuenca Grau and Ian Horrocks, *Privacy-preserving query answering in logic-based information systems*, Proceedings of the 2008 Conference on ECAI 2008: 18th European Conference on Artificial Intelligence (Amsterdam, The Netherlands, The Netherlands), IOS Press, 2008, pp. 40–44.

Bibliografia III



Wilma Maria da Costa Medeiros, *Sisont: Sistema de informação em saúde baseado em ontologias*, Universidade Federal do Rio Grande do Norte, 2009.



Tribunal Superior de Justiça, *Lei Geral de Proteção de Dados Pessoais (LGPD) Lei nº 13.853, de 2019.*, https://www.jusbrasil.com.br/legislacao/612902269/lei-13709-18#art-5_inc-XII, 2019, [Acesso: 15/06/2023].



Cynthia Dwork and Aaron Roth, *The algorithmic foundations of differential privacy.*, Foundations and Trends in Theoretical Computer Science **9** (2014), no. 3-4, 211–407.



Cynthia Dwork and Adam Smith, *Differential privacy for statistics: What we know and what we want to learn*, Journal of Privacy and Confidentiality **1** (2010), no. 2.

Bibliografía IV



Cynthia Dwork, *Differential privacy*, Automata, Languages and Programming (Berlin, Heidelberg) (Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, eds.), Springer Berlin Heidelberg, 2006, pp. 1–12.



Eldora, Martin Knechtel, and Rafael Peñaloza, *Correcting access restrictions to a consequence more flexibly*, Description Logics (Riccardo Rosati, Sebastian Rudolph, and Michael Zakharyashev, eds.), vol. 745, CEUR-WS.org, 2011.



Ulfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova, *Rappor: Randomized aggregatable privacy-preserving ordinal response*, Proceedings of the 21st ACM Conference on Computer and Communications Security (Scottsdale, Arizona), 2014.

Bibliografía V



Bernardo Cuenca Grau and Egor V. Kostylev, *Logical foundations of privacy-preserving publishing of linked data*, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA., 2016, pp. 943–949.



Bernardo Cuenca Grau and Boris Motik, *Reasoning over ontologies with hidden content: The import-by-query approach*, CoRR **abs/1401.5853** (2014).



Nicola Guarino, *Formal ontology and information systems*, IOS Press, 1998, pp. 3–15.



Seiji Isotani and Ig Ibert Bittencourt, *Dados abertos conectados*, Novatec, 2015.

Bibliografía VI



Elisa F. Kendall and Deborah L. McGuinness, *Ontology engineering*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool Publishers, 2019.



Laura Hazard Owen, *Glow: The Best Fertility Tracking App*, <https://thenightlight.com/best-fertility-tracking-app/>, 2016, [Last updated: October, 2017, Online: accessed 26 October, 2020].



Steffen Staab and Rudi Studer (eds.), *Handbook on ontologies*, International Handbooks on Information Systems, Springer, 2004.



Alexandra Wood, Micah Altman, Kobbi Nissim, and Salil Vadhan, *Designing access with differential privacy*, ch. 6, J Pal, 2020.

