

Web ngữ nghĩa

Soạn bởi: Nguyễn Bá Ngọc

Chương 1

Hà Nội-2021


Chương 1.

Tổng quan về Web ngữ nghĩa

Nội dung

- 1.1. Khái niệm Web ngữ nghĩa
- 1.2. Web ngữ nghĩa hiện nay
- 1.3. Đồ thị tri thức
- 1.4. Tổng quan công nghệ Web ngữ nghĩa
- 1.5. Ontology và một số khái niệm quan trọng

Nội dung

- 
- 1.1. Khái niệm Web ngữ nghĩa
 - 1.2. Web ngữ nghĩa hiện nay
 - 1.3. Đồ thị tri thức
 - 1.4. Tổng quan công nghệ Web ngữ nghĩa
 - 1.5. Ontology và một số khái niệm quan trọng

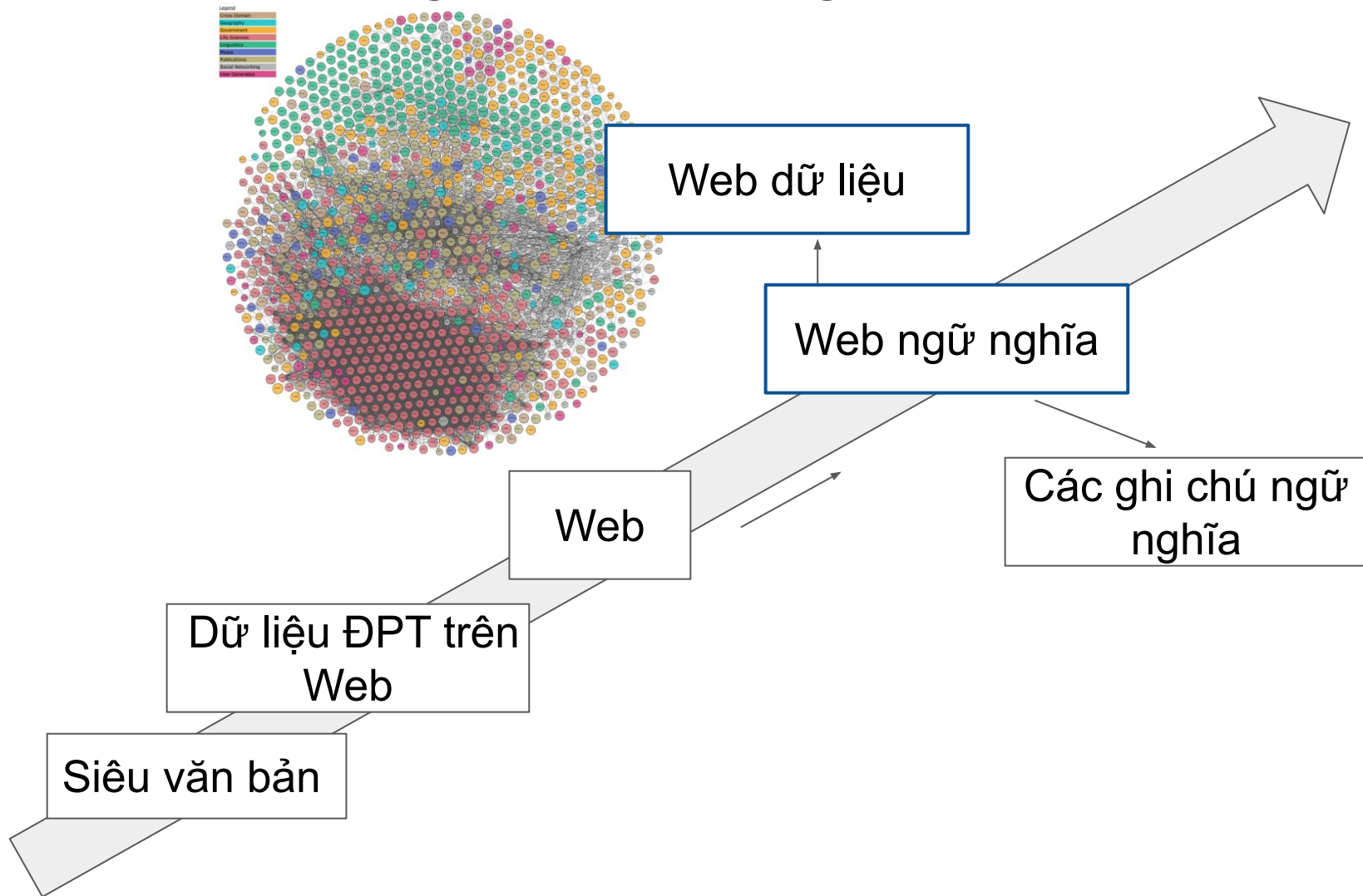
Vấn đề xử lý dữ liệu Web

- Web là nguồn tri thức lớn của nhân loại
- Tuy nhiên có nhiều yếu tố khiến việc khai thác hiệu quả gặp nhiều khó khăn
 - Được thiết kế cho người, không phải cho máy
 - Các nội dung chủ yếu được cung cấp dưới dạng văn bản, âm thanh, hình ảnh, video
 - Phù hợp với nhu cầu sử dụng của người, tuy nhiên khó xử lý ngữ nghĩa bằng máy tính...
 - Khó phát triển các ứng dụng để cung cấp thông tin hữu ích cho người dùng
 - ... Khả năng xử lý ngữ nghĩa bằng máy tính là tiềm năng lớn để phát triển nhiều ứng dụng hữu ích.

Vấn đề xử lý dữ liệu Web₍₂₎

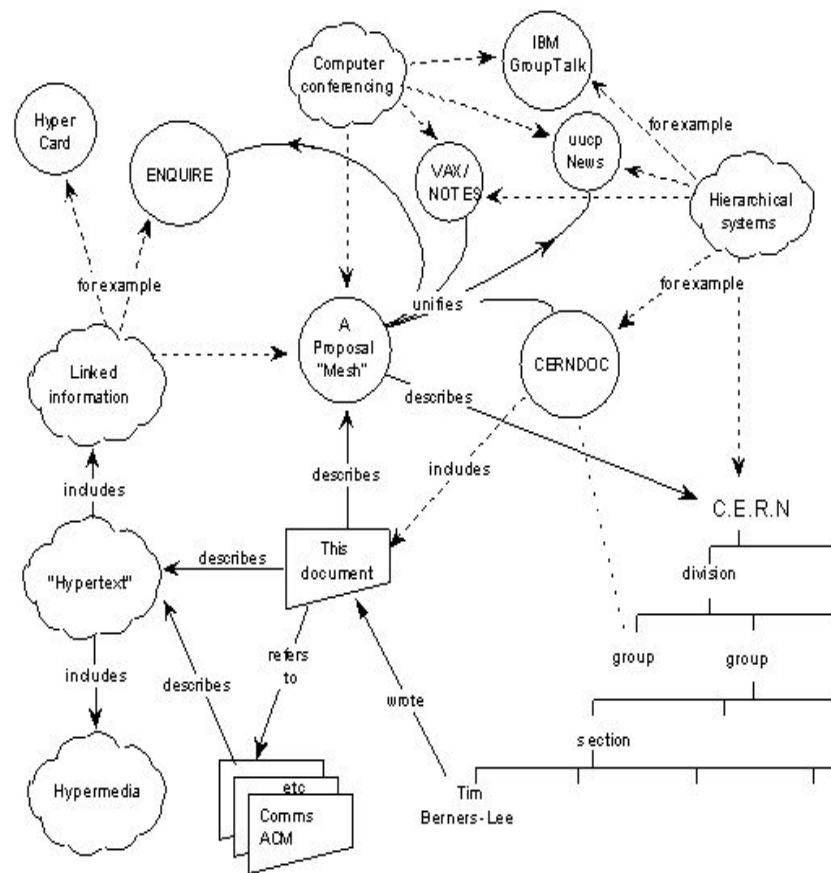
- Truy cập nội dung Web được thực hiện chủ yếu thông qua cơ chế tìm kiếm thông tin
 - *(Với các công nghệ hiện tại)*
 - Các truy vấn dựa trên từ khóa (ngôn ngữ tự nhiên);
 - Kết quả là danh sách các tài liệu được xếp hạng;
 - Người dùng tự rà soát các tài liệu để lấy thông tin phù hợp.
- Trong nhiều trường hợp, người dùng muốn có những kết quả trực tiếp:
 - Ví dụ, giá trị số của số PI? => 3.14159265359
 - => Cần khả năng xử lý ngữ nghĩa bằng máy tính và dữ liệu có cấu trúc.

Dữ liệu trong môi trường Web



Sự khởi đầu của Web và Web ngữ nghĩa

- Sự khởi đầu của Web:
 - Năm 1989 Tim Berners-Lee đã mô tả mạng lưới tài nguyên liên kết như sự hợp nhất của nhiều hoạt động quản lý thông tin, khởi đầu cho sự phát triển của Web.
- Sự hình thành Web ngữ nghĩa:
 - Guha MCF (~94)
 - XML+MCF=>RDF(~96)
 - RDF+OO=>RDFS(~99)
 - RDFS+KR=>DAML+OIL(00)
 - Các hoạt động của W3C (01)
 - Quy chuẩn OWL của W3C (03)
 - ...



<http://www.w3.org/History/1989/proposal.html>

Các khái niệm Web ngữ nghĩa

- Web ngữ nghĩa 1

- Khái niệm đầu tiên được đưa ra bởi Tim Berners-Lee, ước mơ về Web của thông tin mà máy tính có thể hiểu
- Không giới hạn công nghệ được sử dụng
 - Có thể áp dụng các phương pháp học thống kê (NLP, học máy, v.v..)
- Người dùng cuối là con người

"Web ngữ nghĩa là 1 mở rộng của Web hiện tại trong đó ý nghĩa của thông tin được thiết lập tường minh, mở ra khả năng cộng tác tốt hơn giữa người và máy tính." [Berners-Lee, Hendler and Lassila, The Semantic Web, Scientific American, 2001]

Máy tính cũng có thể hiểu thông tin trong môi trường Web ngữ nghĩa



Các khái niệm Web ngữ nghĩa₍₂₎

- Web ngữ nghĩa 2
 - Được định hướng bởi W3C: Web của dữ liệu
 - Sử dụng các quy chuẩn, ví dụ: RDF(S), OWL, SPARQL, SHACL, v.v.
 - Cho phép lưu và chia sẻ dữ liệu trên Web, tạo các bộ từ vựng và các luật xử lý dữ liệu.
 - Hỗ trợ tương tác máy-máy làm nền tảng cho các ứng dụng phục vụ con người

Vì sao khó xử lý ngữ nghĩa trong HTML?

HTML

Ngữ nghĩa

<h1>

As We May Think

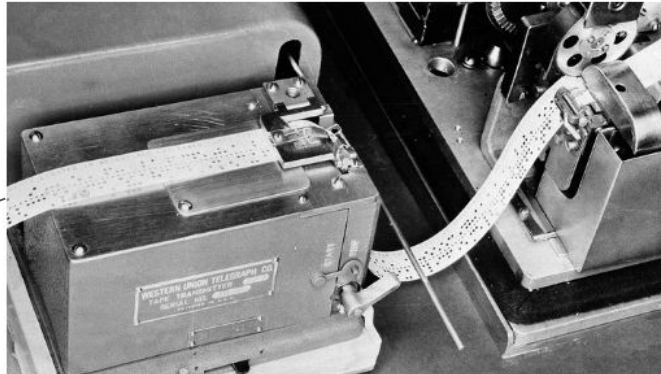
"Consider a future device ... in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Tiêu đề bài viết

VANNEVAR BUSH JULY 1945 ISSUE

Tác giả & năm
xuất bản

<picture>



Hình minh họa máy
điện tín

The telegram was a breakthrough in communication technology, which Vannevar Bush imagined could evolve in unprecedented ways. (AP)

<p><i>

As Director of the Office of Scientific Research and Development, Dr. Vannevar Bush has coordinated the activities of some six thousand leading American scientists in the application of science to warfare. In this significant article he holds up an incentive for scientists when the fighting has ceased. He urges that men of science should then turn to the massive task of making more accessible our bewildering store of knowledge. For years inventions have extended man's physical powers rather than the powers of his mind. Trip hammers that multiply the fists, microscopes that sharpen the eye, and engines of destruction and detection are new results, but not the end results, of modern science. Now, says Dr. Bush, instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages. The perfection of these pacific instruments should be the first objective of our scientists as they emerge from their war work. Like Emerson's famous address of 1837 on "The American Scholar," this paper by Dr. Bush calls for a new relationship between thinking man and the sum of our knowledge. — THE EDITOR

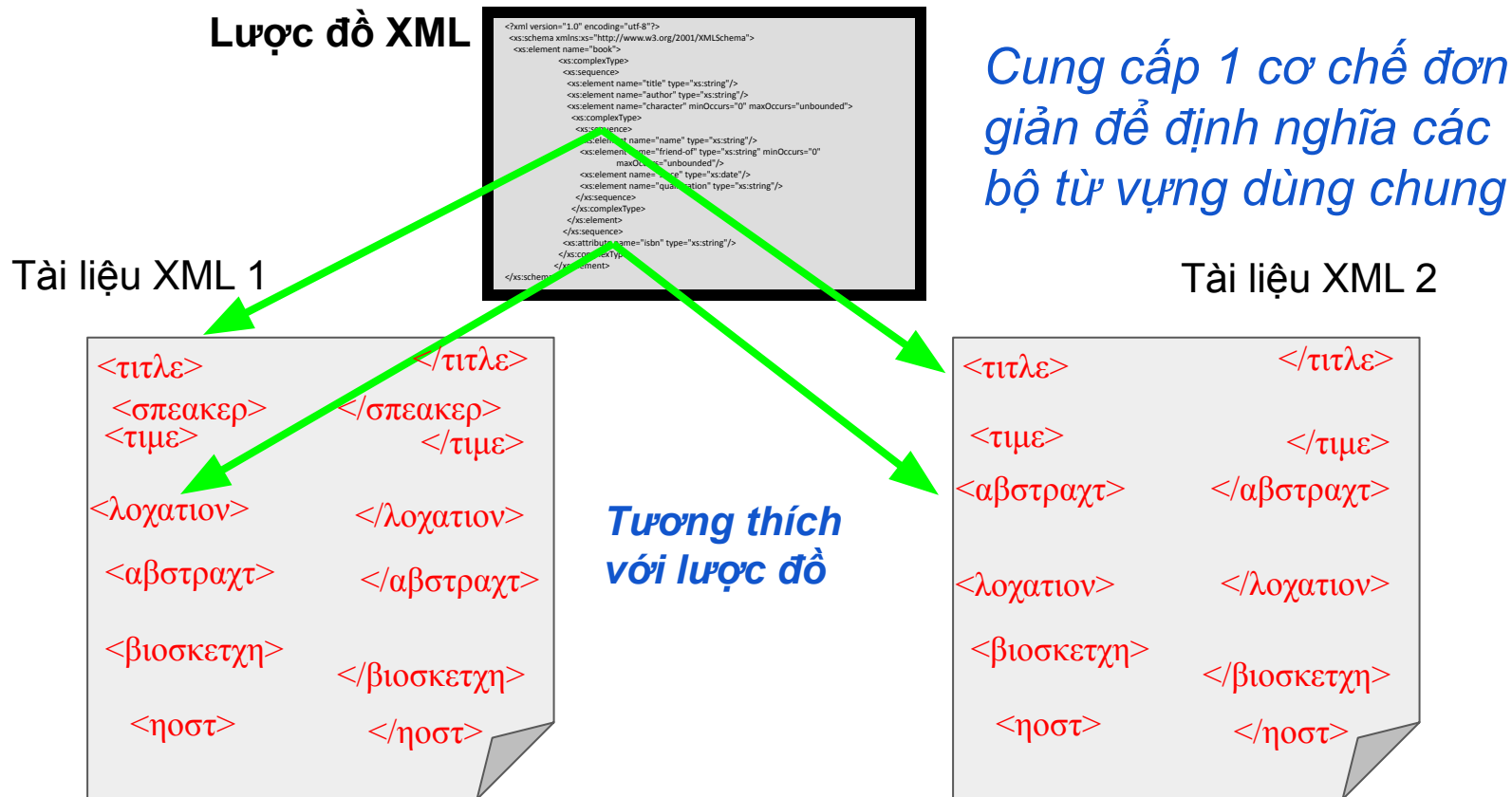
Lời giới thiệu của
người biên soạn

Các thẻ chủ yếu gợi ý cách hiển thị

Có thể giải quyết các vấn đề bằng XML?

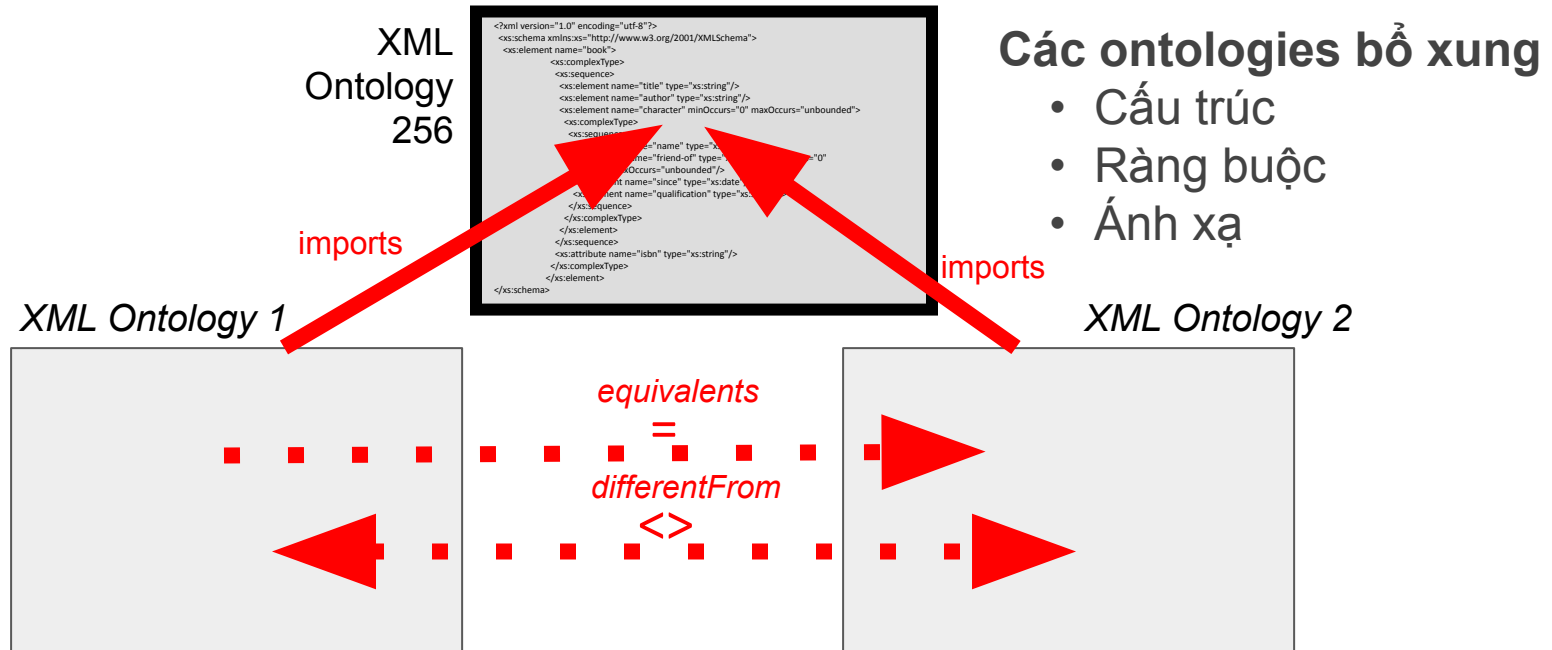
- Người dùng XML tự tạo tập thẻ riêng cho từng ứng dụng
 - Có thể dùng các thẻ <title>, <author>, v.v..
- XML ≠ biểu diễn ngữ nghĩa cho máy
 - Các tên thẻ không hàm chứa ngữ nghĩa hình thức để có thể xử lý bằng máy tính.
 - DTD hàm chứa rất ít hoặc không hàm chứa ngữ nghĩa.
 - Mô tả ràng buộc cấu trúc của các phần tử.
 - Lược đồ XML cung cấp 1 cơ chế đơn giản để định nghĩa các bộ từ vựng dùng chung.
 - **Hỗ trợ** giải quyết vấn đề xử lý ngữ nghĩa ở 1 mức độ hạn chế.

Các lược đồ XML



Có thể dẫn tới quá nhiều lược đồ XML... Nhưng không có cơ chế liên kết các lược đồ XML

Cần phát triển ngôn ngữ biểu diễn tri thức



Có thể được biểu diễn bằng XML và cho phép:
Liên kết các bộ từ vựng;
Máy tính có thể hiểu ngữ nghĩa (ở 1 mức độ nhất định).

Nội dung

1.1. Khái niệm Web ngữ nghĩa

1.2. Web ngữ nghĩa hiện nay

1.3. Đồ thị tri thức

1.4. Tổng quan công nghệ Web ngữ nghĩa

1.5. Ontology và một số khái niệm quan trọng

Trạng thái hiện tại

- Phiên bản Web ngữ nghĩa của W3C đã được phát triển liên tục
- Các ngôn ngữ và quy chuẩn Web ngữ nghĩa được sử dụng trong nhiều hệ thống và dịch vụ Web:
 - Google và Facebook phát hiện và sử dụng (một số) dữ liệu RDF ở dạng nhúng trong các trang HTML
 - Google, Yahoo, Microsoft và Yandex phát triển những bộ từ vựng dùng chung (schema.org)
 - lod-cloud.net có nhiều bộ dữ liệu mở ở định dạng RDF
 - v.v..

Các dịch vụ Web hiện có

- Google có thể là 1 ví dụ tiêu biểu, tuy nhiên các công ty khác cũng có những giải pháp tương đương:
 - 2010 Google thu tóm MediaWeb và cơ sở tri thức Freebase của nó;
 - 2014: Freebase có 1.2 tỷ dữ kiện về khoảng 43 triệu thực thể;
 - 2015+: Đồ thị tri thức của Google, được cập nhật tự động bằng các kỹ thuật trích rút thông tin.

https://vi.wikipedia.org/wiki/Trịnh_Công_Sơn · [Translate this page](#)

Trịnh Công Sơn – Wikipedia tiếng Việt

Trịnh Công Sơn (28 tháng 2 năm 1939 – 1 tháng 4 năm 2001) là một nam nhạc sĩ người Việt Nam. Ông được coi là một trong những nhạc sĩ lớn nhất của Tân nhạc ...

Ca sĩ trình bày thành công: [Khánh Ly](#); [Hồng Nhung](#)... Ca khúc tiêu biểu: [Diễm xưa](#); [Biển nhớ](#); [Cát...](#)
 Năm hoạt động: 1958–2001 Nghề nghiệp: [Nhạc sĩ](#); [Ca sĩ](#); [Nhạc công](#); [H...](#)

[Danh sách bài hát của Trịnh...](#) · [Nhạc phản chiến của Trịnh...](#) · [Khánh Ly](#) · [Diễm xưa](#)

Videos >



Nhạc Trịnh Công Sơn Chọn Lọc Hay Nhất Đi Cùng Năm ...

YouTube · Nhạc Trịnh Công Sơn
Jul 14, 2022



48 Tình Khúc Nhạc Trịnh Công Sơn Hay Nhất Mọi Thời Đại ...

YouTube · Nhạc Trịnh Công Sơn
Jul 19, 2022



28 Tình Khúc Nhạc Trịnh Công Sơn Hay Nhất Mọi Thời Đại ...

YouTube · Nhạc Trịnh Công Sơn
Jul 12, 2022



Nhạc Trịnh Công Sơn Chọn Lọc Hay Nhất Hiếm Có Khó Tìm ...

YouTube · Nhạc Trịnh Công Sơn
Jun 24, 2022

[View all](#) →

Listen



YouTube



Spotify



Apple Music



Deezer

About

Trịnh Công Sơn was a Vietnamese, musician, songwriter, painter and poet. He is widely considered to be Vietnam's best songwriter. [Wikipedia](#)

Born: February 28, 1939, [Buon Ma Thuot](#)

Died: April 1, 2001, [Ho Chi Minh City](#)

Buried: April 4, 2001, [Quang Binh Pagoda Cemetery](#)

Siblings: [Trinh Vinh Trinh](#)

[Feedback](#)

Songs

[Ru ta ngậm ngùi](#)

[Ru em](#) · 1997



[Hạ trắng](#)

[Ru ta ngậm ngùi](#) - Latin Jazz



[Tuổi đá buồn](#)

[Ru em](#) · 1997



[Biển nhớ](#)

[Hòa tấu Vọng 2 - Hòa tấu Guitar](#) · 1996



Facebook

The Open Graph protocol



Introduction

The [Open Graph protocol](#) enables any web page to become a rich object in a social graph. For instance, this is used on Facebook to allow any web page to have the same functionality as any other object on Facebook.

While many different technologies and schemas exist and could be combined together, there isn't a single technology which provides enough information to richly represent any web page within the social graph. The Open Graph protocol builds on these existing technologies and gives developers one thing to implement. Developer simplicity is a key goal of the Open Graph protocol which has informed many of the [technical design decisions](#).

Basic Metadata

To turn your web pages into graph objects, you need to add basic metadata to your page. We've based the initial version of the protocol on [RDFa](#) which means that you'll place additional `<meta>` tags in the `<head>` of your web page. The four required properties for every page are:

- `og:title` - The title of your object as it should appear within the graph, e.g., "The Rock".
- `og:type` - The [type](#) of your object, e.g., "video.movie". Depending on the type you specify, other properties may also be required.
- `og:image` - An image URL which should represent your object within the graph.
- `og:url` - The canonical URL of your object that will be used as its permanent ID in the graph, e.g., "https://www.imdb.com/title/tt0117500/".

As an example, the following is the Open Graph protocol markup for [The Rock on IMDB](#):

```
<html prefix="og: https://ogp.me/ns#">
<head>
<title>The Rock (1996)</title>
<meta property="og:title" content="The Rock" />
```

Apple

Siri sử dụng đồ thị tri thức để hiểu yêu cầu của người dùng

Trợ lý ảo

Calls and Texts

Siri lets you stay connected without lifting a finger.

Siri can make calls or send texts for you whether you are driving, have your hands full, or are simply on the go.¹ It can even announce your messages on your AirPods.² It also offers proactive suggestions — like texting someone that you're running late for a meeting — so you can stay in touch effortlessly.³

"Hey Siri, call Mom on speaker"



Siri reminds you to make the calls that matter.



"Text Donna 'I'm on the way exclamation point'"

Các dữ kiện cho cơ sở tri thức từ đâu tới?

- Các CSTT như DBpedia và Freebase đã khởi đầu bằng cách biểu diễn dữ liệu trong Wikipedia bằng các lược đồ riêng.
- Các công nghệ Web ngữ nghĩa giống như mã nguồn mở trong biểu diễn tri thức.
 - Vẫn đang liên tục được phát triển.
- Dữ liệu nhúng trong trang Web.
- Trích xuất dữ liệu từ các tài liệu văn bản phi cấu trúc, ví dụ: Các bài viết, bài báo, v.v.

Cơ sở tri thức mở: DBpedia

- DBpedia là 1 cơ sở tri thức mở dựa trên các công nghệ Web ngữ nghĩa:
 - 800 triệu dữ kiện về khoảng 4.6 triệu thực thể từ Wikipedia tiếng Anh, có dữ liệu trong khoảng 21 ngôn ngữ;
 - Hỗ trợ tích hợp 90 tỉ dữ kiện từ hơn 1000 tập dữ liệu RDF trong không gian dữ liệu liên kết mở.

DBPedia

*Dữ liệu từ
Wikipedia trong
RDF*

SPARQL Explorer for <http://dbpedia.org/sparql>

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT distinct ?soccerplayer ?countryOfBirth ?team ?countryOfTeam ?stadiumcapacity
{
  ?soccerplayer a dbo:SoccerPlayer ;
    dbo:position|dbp:position <http://dbpedia.org/resource/Goalkeeper_(association_football)> ;
    dbo:birthPlace|dbo:country* ?countryOfBirth ;
    #dbo:number 13 ;
    dbo:team ?team .
    ?team dbo:capacity ?stadiumcapacity ; dbo:ground ?countryOfTeam .
    ?countryOfBirth a dbo:Country ; dbo:populationTotal ?population .
    ?countryOfTeam a dbo:Country .
  FILTER (?countryOfTeam != ?countryOfBirth)
  FILTER (?stadiumcapacity > 30000)
  FILTER (?population > 10000000)
  order by ?soccerplayer
```

Results:

SPARQL results:

soccerplayer	countryOfBirth	team	countryOfTeam	stadiumcapacity
:Abdesslam_Benabdellah ↗	:Algeria ↗	:Wydad_Casablanca ↗	:Morocco ↗	67000
:Ailton_Moraes_Michellon ↗	:Brazil ↗	:FC_Red_Bull_Salzburg ↗	:Austria ↗	31000
:Alain_Gouaméné ↗	:Ivory_Coast ↗	:Raja_Casablanca ↗	:Morocco ↗	67000
:Allan_McGregor ↗	:United_Kingdom ↗	:Beşiktaş_J.K. ↗	:Turkey ↗	41903
:Anthony_Scribe ↗	:France ↗	:FC_Dinamo_Tbilisi ↗	:Georgia_(country) ↗	54549
:Brahim_Zaari ↗	:Netherlands ↗	:Raja_Casablanca ↗	:Morocco ↗	67000
:Bréiner_Castillo ↗	:Colombia ↗	:Deportivo_Táchira ↗	:Venezuela ↗	38755
:Carlos_Luis_Morales ↗	:Ecuador ↗	:Club_Atlético_Independiente ↗	:Argentina ↗	48069
:Carlos_Navarro_Montoya ↗	:Colombia ↗	:Club_Atlético_Independiente ↗	:Argentina ↗	48069
:Cristián_Muñoz ↗	:Argentina ↗	:Colo-Colo ↗	:Chile ↗	47000
:Daniel_Ferreira ↗	:Argentina ↗	:FBC_Melgar ↗	:Peru ↗	60000
:David_Bíčík ↗	:Czech_Republic ↗	:Karşıyaka_S.K. ↗	:Turkey ↗	51295
:David_Loria ↗	:Kazakhstan ↗	:Karşıyaka_S.K. ↗	:Turkey ↗	51295
:Denys_Boyko ↗	:Ukraine ↗	:Beşiktaş_J.K. ↗	:Turkey ↗	41903
:Eddie_Gustafsson ↗	:United_States ↗	:FC_Red_Bull_Salzburg ↗	:Austria ↗	31000
:Emilian_Dolha ↗	:Romania ↗	:Lech_Poznań ↗	:Poland ↗	43269
:Eusebio_Acasuzo ↗	:Peru ↗	:Club_Bolívar ↗	:Bolivia ↗	42000
:Faryd_Mondragón ↗	:Colombia ↗	:Real_Zaragoza ↗	:Spain ↗	34596
:Faryd_Mondragón ↗	:Colombia ↗	:Club_Atlético_Independiente ↗	:Argentina ↗	48069
:Federico_Vilar ↗	:Argentina ↗	:Club_Atlas ↗	:Mexico ↗	54500
:Fernando_Martinuzzi ↗	:Argentina ↗	:Real_Garcilaso ↗	:Peru ↗	45000
:Fábio André da Silva ↗	:Portugal ↗	:Servette_FC ↗	:Switzerland ↗	30084

Wikidata

- + Hướng tới tạo 1 CSTT có thể đọc và cập nhật bởi người và máy.
- + Giữ vai trò trung tâm lưu trữ dữ liệu có cấu trúc cho các bộ Bách khoa toàn thư khác như Wikipedia, Wikivoyage, Wiktionary, Wikisource, v.v..



The screenshot shows the Wikidata Main Page. At the top is the Wikidata logo and a navigation bar with links: Main Page, Discussion, Read, View source, View history, and a search bar. Below the navigation bar is a large central box with the text "Welcome to Wikidata" and "the free knowledge base with 91,786,616 data items that anyone can edit." Below this box are links for Introduction, Project Chat, Community Portal, and Help. To the left of the main content is a sidebar with various links and tools. To the right of the main content are three boxes: "Welcome!", "Learn about data", and "Get involved".

Wikidata

Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item

In other projects
Wikimedia Commons
MediaWiki
Meta-Wiki
Wikispecies
Wikibooks
Wikimania
Wikinews
Wikipedia
Wikiquote
Wikisource
Wikiversity
Wikivoyage
Wiktionary

In Wikipedia
العربية

Main Page Discussion Read View source View history Search Wikidata

collab

open

Welcome to Wikidata

the free knowledge base with 91,786,616 data items that anyone can edit.

free

Introduction • Project Chat • Community Portal • Help

mi

Want to help translate? Translate the missing messages.

Welcome!

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a [free license](#), exported using [standard formats](#), and can be [interlinked to other open data sets](#) on the linked data web.

Learn about data

New to the wonderful world of data? Develop and improve your data literacy through content designed to get you up to speed and feeling comfortable with the fundamentals in no time.

Item: *Earth* (Q2) Property: *highest point* (P610)



custom value:

Get involved

For a complete starters' guide, visit the [community portal](#).

[Learn about Wikidata](#)

Ví dụ 1.1. Truy vấn trên Wikidata

Country populations together with total city populations

Map

218 results in 4388 ms

Code

Download

Link

Search

country	countryLabel	population	totalCityPopulation	ratio
Q148	People's Republic of China	1409517397	115388039	12.21545499182978575448
Q668	India	1326093247	147242392	9.00619195998934871963
Q30	United States of America	328239523	10936541	30.0131022230886346972
Q15180	Soviet Union	293047571	28500	10282.37091228070175438596
Q252	Indonesia	270625568	10961704	24.6882754724995310948
Q843	Pakistan	216565318	43101219	5.02457524461199113649
Q155	Brazil	210147125	90381055	2.3251236113585972193
Q1033	Nigeria	190886311	44761481	4.26452178827595092307
Q34266	Russian Empire	181537800	69825	2599.89688506981740064447
Q34266	Russian Empire	178378800	69825	2554.65520945220193340494
Q902	Bangladesh	164669751	22627219	7.27750728005947173623
Q159	Russia	146804372	1135276	129.31161409207981142912
Q96	Mexico	130526945	34239370.789	3.81218877544134299716
Q17	Japan	126434565	23318712	5.42202180806555696558
Q34266	Russian Empire	125640021	69825	1799.35583243823845327605
Q7318	Nazi Germany	109518183	7741	14147.80816431985531585067
Q115	Ethiopia	104957438	2612647	40.17283544236936715905
Q79	Egypt	94798827	28436468	3.33370610583564738068
Q881	Vietnam	94660000	13370442	7.07979586613516591299
Q974	Democratic Republic of the Congo	86790567	17871025	4.85649631176723215372

Nội dung

1.1. Khái niệm Web ngữ nghĩa

1.2. Web ngữ nghĩa hiện nay

1.3. Đồ thị tri thức

1.4. Tổng quan công nghệ Web ngữ nghĩa

1.5. Ontology và một số khái niệm quan trọng

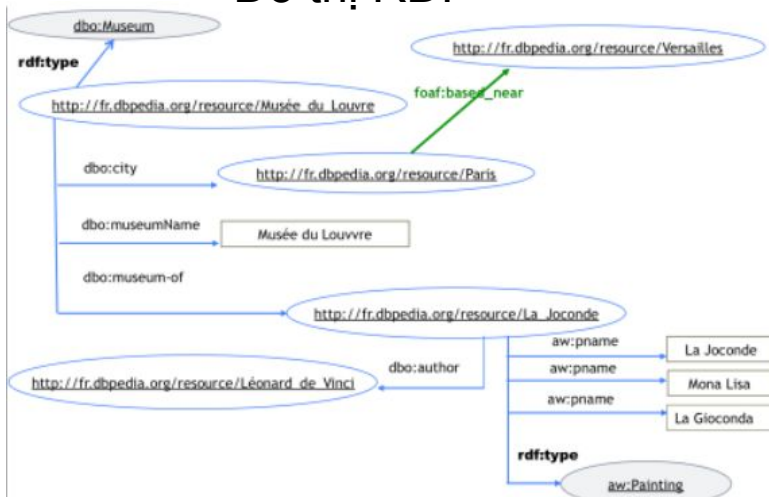
Khái niệm đồ thị tri thức

- Đồ thị tri thức là 1 mạng quy mô lớn của các thực thể, kiểu có nghĩa của chúng, thuộc tính, và mối quan hệ giữa các thực thể. -> Đồ thị RDF
 - "Knowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities" [Journal of Web Semantics]
- Đồ thị tri thức có thể được biểu diễn như đồ thị RDF, bao gồm tập các bộ-3 RDF.
 - [Farber et al.]

Đồ thị tri thức ~ Cơ sở tri thức dạng đồ thị.

Đồ thị tri thức (KG)

Đồ thị RDF



Bộ từ vựng



Truy vấn (SPARQL)

```
PREFIX dbo: <http://dbpedia.org/ontology#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?m ?p
WHERE { ?m rdf:type dbo:Museum . ?m dbo:museum-of ?p }
```

Suy diễn: (Pellet, Fact++, Hermit, v.v..)

- + Mở rộng CSTT: Bổ sung thêm những dữ kiện có thể được suy ra từ những gì đang có.
- + Kiểm tra tính nhất quán: Không có mâu thuẫn
- + Sửa lỗi CSTT
- + v.v..

Các mệnh đề và luật

```
owl:equivalentClass(dbo:Municipality, dbo:Place)
owl:equivalentClass(dbo:Place, dbo:Wikidata:Q532)
owl:equivalentClass(dbo:Village, dbo:PopulatedPlace)
owl:equivalentClass(dbo:PopulatedPlace, dbo:Municipality)
owl:disjointClass(dbo:PopulatedPlace, dbo:Artist)
owl:disjointClass(dbo:PopulatedPlace, dbo:Painting)
owl:FunctionalProperty(dbo:city)
owl:InverseFunctionalProperty(dbo:museum-of)
```

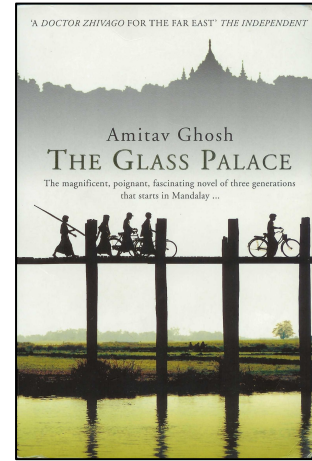
```
dbo:birthPlace(X, Y) => dbo:citizenOf(X, Y)
dbo:parentOf(X, Y) => dbo:child(Y, X)
```

Quản lý đồ thị tri thức

- Mở rộng: các đồ thị tri thức luôn cần được cập nhật
 - Liên kết dữ liệu (phát hiện thực thể, các trùng lặp, ...).
 - Dự đoán liên kết: Thêm mới liên kết.
 - Kết nối ontology: Kết nối các đồ thị.
 - Dự đoán/suy diễn các dữ kiện mới.
- Kiểm tra: Đồ thị tri thức có thể chứa lỗi
 - Kiểm tra liên kết.
 - Xử lý lỗi và các vấn đề đa nghĩa.
- Suy diễn: Liệu có thể phát hiện tri thức mới?
 - Phát hiện tri thức.
 - Suy diễn và lập kế hoạch tự động.
- V.V..

Ví dụ 1.2. Tích hợp dữ liệu

Rất quan trọng đối với dữ liệu Web do có nhiều nguồn phân tán



Sách

ID	Author	Title	Publisher	Year
ISBN 0-00-6511409-X	id_xyz	The Glass Palace	id_qpr	2000

Tác giả

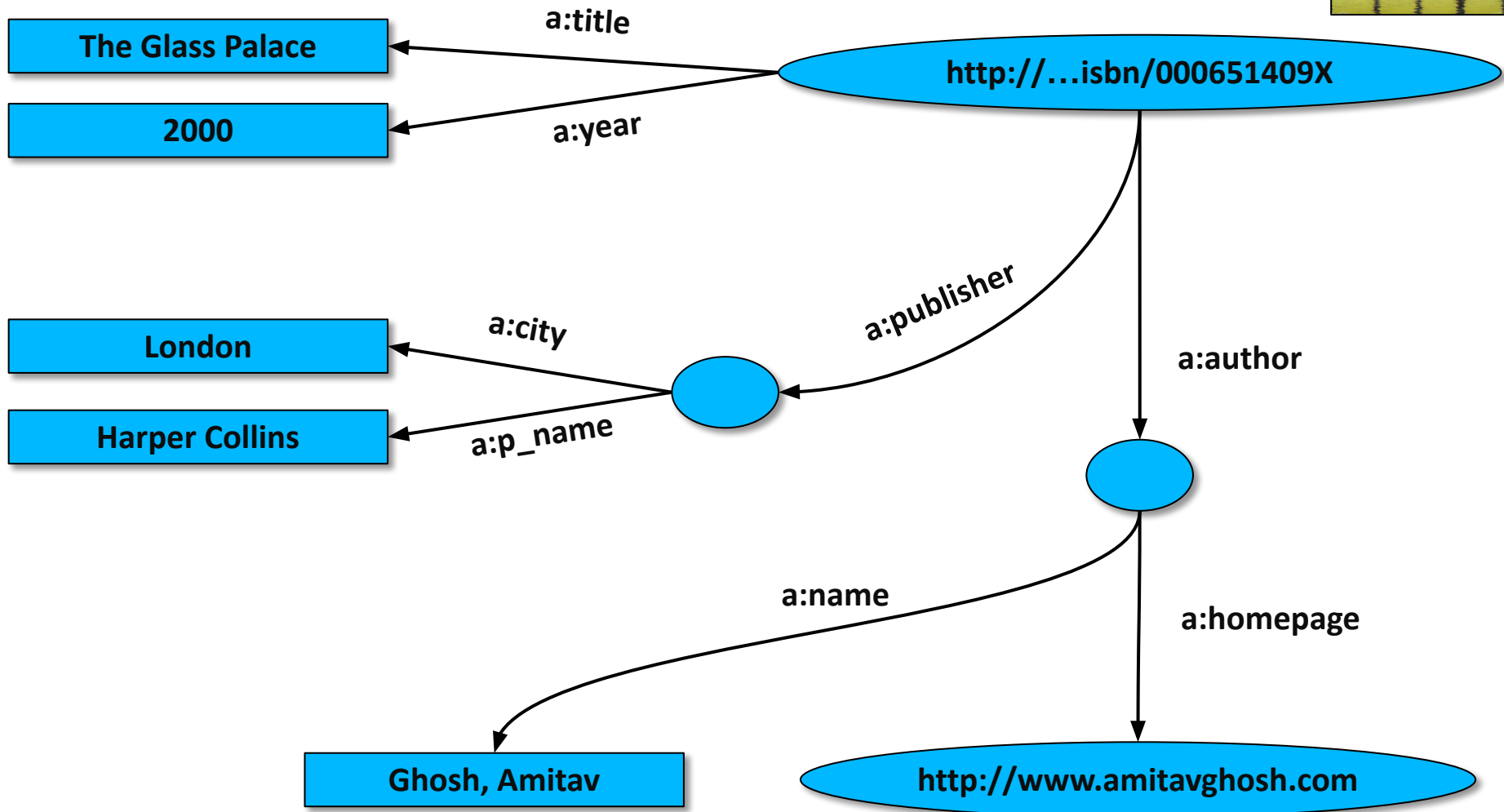
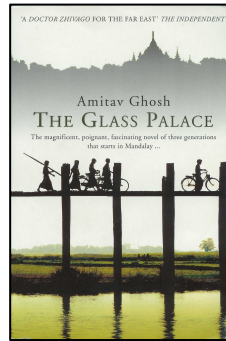
ID	Name	Homepage
id_xyz	Ghosh, Amitav	http://www.amitavghosh.com

Nhà xuất bản

ID	Publisher's name	City
id_qpr	Harper Collins	London

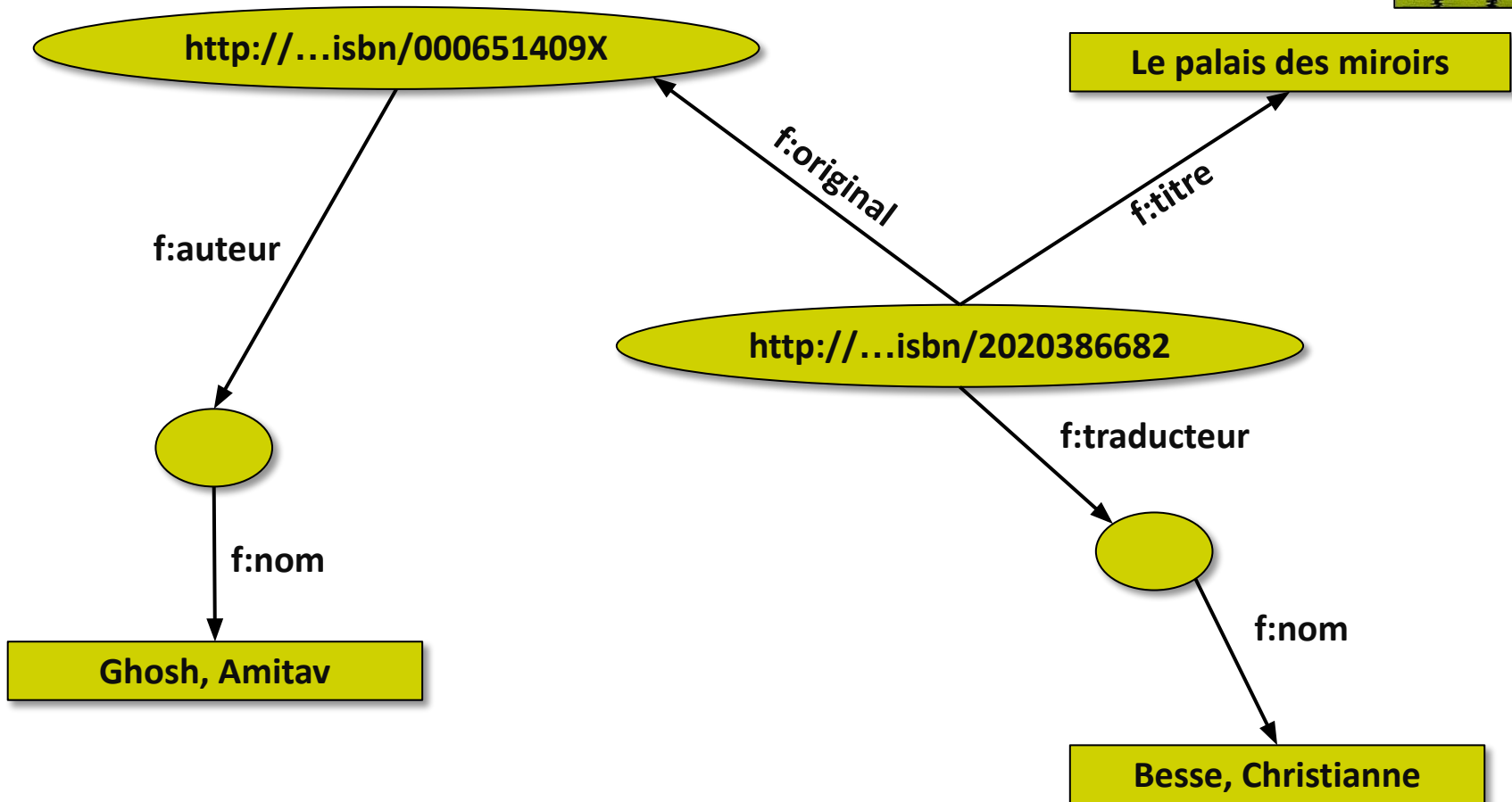
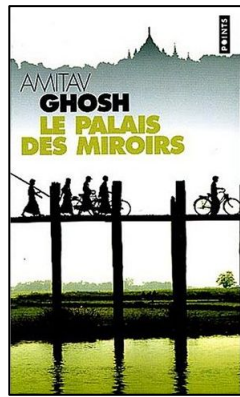
Ví dụ 1.2. Tích hợp dữ liệu₍₂₎

Biểu diễn bằng đồ thị RDF

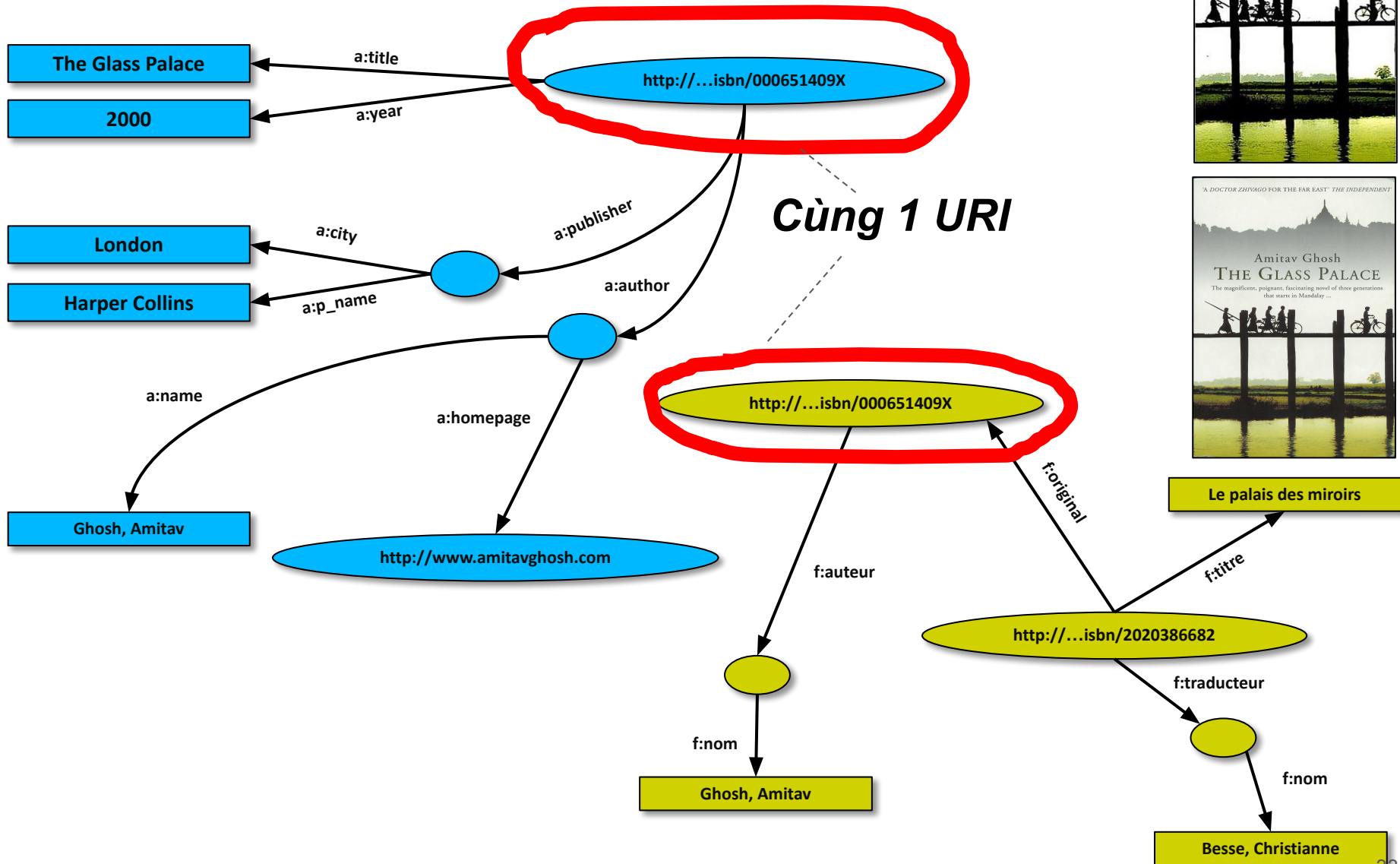


Ví dụ 1.2. Tích hợp dữ liệu⁽³⁾

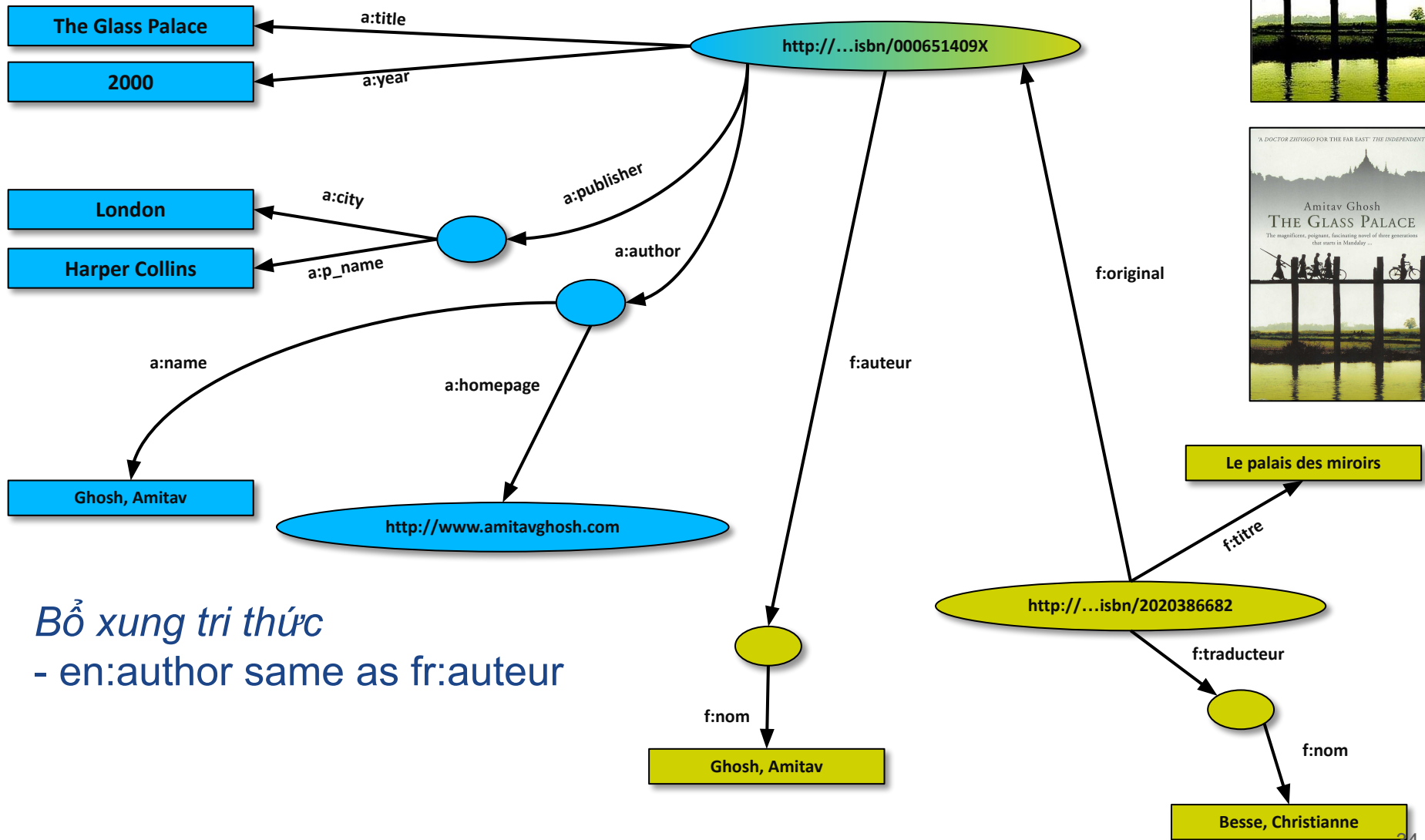
Phiên bản tiếng Pháp



Ví dụ 1.2. Tích hợp dữ liệu⁽⁴⁾



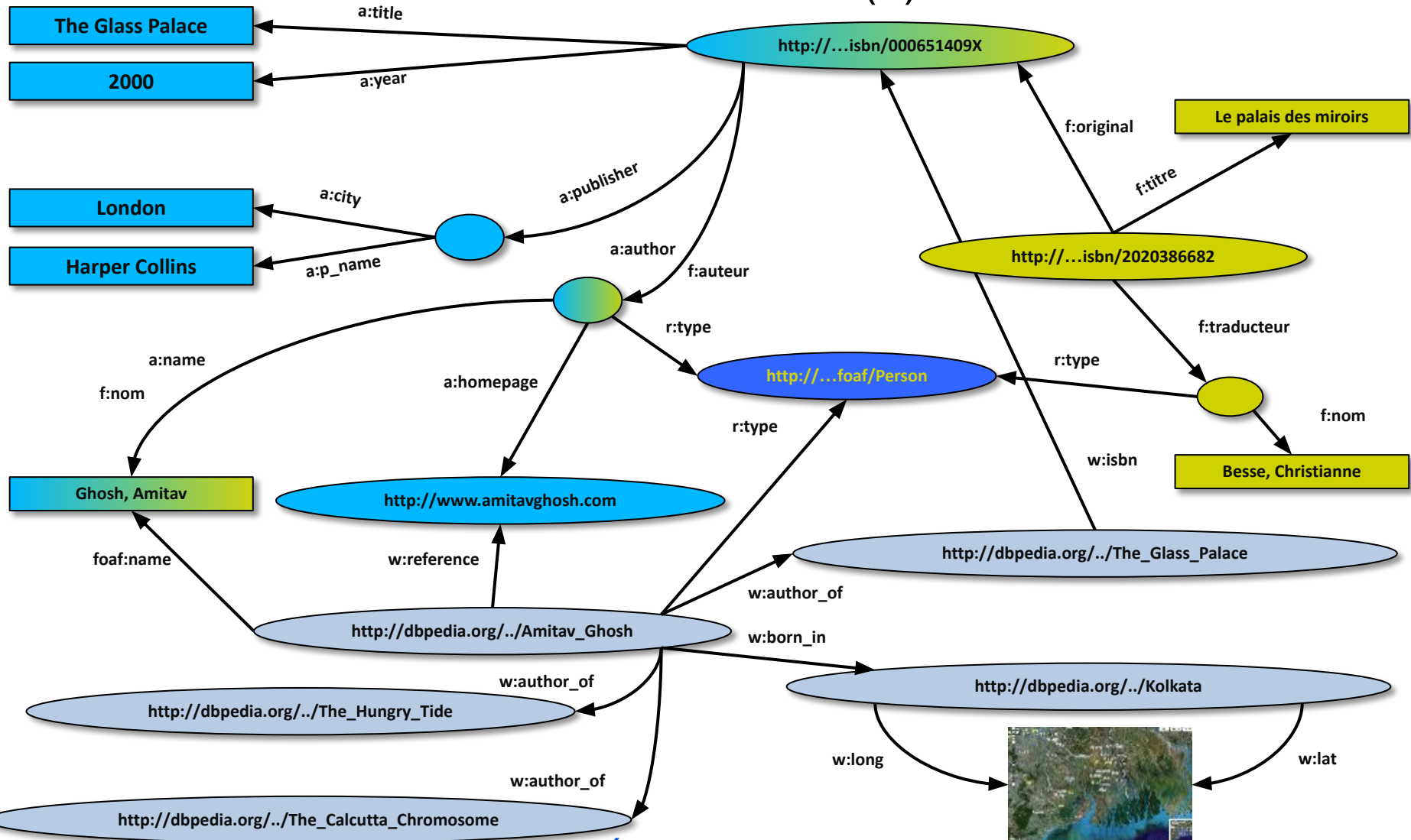
Ví dụ 1.2. Tích hợp dữ liệu⁽⁵⁾



Bổ xung tri thức

- en:author same as fr:auteur

Ví dụ 1.2. Tích hợp dữ liệu⁽⁶⁾



Tiếp tục tích hợp với DBPedia
Làm sao để xác định các lớp, thuộc tính, thực thể tương đương?

Web ngữ nghĩa & đồ thị tri thức

- Web ngữ nghĩa cung cấp nền tảng công nghệ cho phép biểu diễn và tích hợp đồ thị tri thức
 - Các đồ thị tri thức quy mô rất lớn, ở quy mô Web
 - Ví dụ, Wikidata, DBpedia, v.v..
- Biểu diễn tri thức bằng ontologies, luật, v.v..
- Cho phép thực hiện nhiều truy vấn phức tạp và hữu ích.

Nội dung

1.1. Khái niệm Web ngữ nghĩa

1.2. Web ngữ nghĩa hiện nay

1.3. Đồ thị tri thức

1.4. Tổng quan công nghệ Web ngữ nghĩa

1.5. Ontology và một số khái niệm quan trọng

Các ngôn ngữ biểu diễn tri thức

- Biểu diễn tri thức và suy diễn (KR&R) luôn là 1 phần quan trọng của TTNT và nhiều lĩnh vực khác
- Nhiều giải pháp đã được xây dựng, triển khai và phát triển từ những năm 1960
- Hầu hết các giải pháp là riêng tư, chỉ được sử dụng bởi người phát triển
- Bắt đầu từ những năm 1990, ngôn ngữ biểu diễn tri thức dùng chung để hỗ trợ tái sử dụng tri thức và các hệ cơ sở tri thức phân tán được quan tâm phát triển
- Các ngôn ngữ Web ngữ nghĩa (ví dụ, OWL) hiện là sự kết tinh của những ý tưởng này
 - Được sử dụng phổ biến và nhất quán giữa nhiều giải pháp.

Cơ sở dữ liệu và Cơ sở tri thức

- Cơ sở dữ liệu (CSDL) vs. cơ sở tri thức (CSTT)
 - CSDL thường có lược đồ dữ liệu (tri thức đơn giản) và rất nhiều dữ liệu;
 - CSTT có thể có lược đồ phức tạp (các ontologies) và các dữ kiện.
- CSTT hỗ trợ suy diễn, ví dụ,

$\text{parent}(\text{?x}, \text{?y}) \Rightarrow$

$\text{person}(\text{?x}), \text{person}(\text{?y}), \text{child}(\text{?y}, \text{?x}), \text{older}(\text{?x}, \text{?y}), \text{?x} \neq \text{?y}$

$\text{Parent}(\text{john}, \text{mary}) \Rightarrow \text{person}(\text{john}), \text{person}(\text{mary}),$
 $\text{child}(\text{mary}, \text{john}), \text{older}(\text{john}, \text{mary}), \text{john} \neq \text{mary}$

Các giải pháp biểu diễn tri thức

- Ngôn ngữ tự nhiên.
- Mã nguồn lập trình.
- Quan hệ vs. Đồ thị vs. Đối tượng.
- Lô-gic vs. Luật vs. Tiến trình.
- Mạng nơ-ron.
- Tenxơ.

Sử dụng nhiều nền tảng biểu diễn tri thức khác nhau có thể dẫn đến những hệ quả gì?

Các giải pháp biểu diễn tri thức₍₂₎

- Lô-gic tính toán là 1 lựa chọn phổ biến
 - $\text{man(NVA)}, \forall x \text{man}(x) \Rightarrow \text{mortal}(x)$
 - Lô-gic có giới hạn: Các dữ kiện, các quan hệ, và "luật" chỉ có thể là đúng (true) hoặc sai (false)
- \Rightarrow Có thể cần biểu diễn và suy diễn với xác suất hoặc các dữ kiện mờ hoặc tri thức.
- \Rightarrow Có thể cần xử lý các dữ kiện, các luật thay đổi theo thời gian.

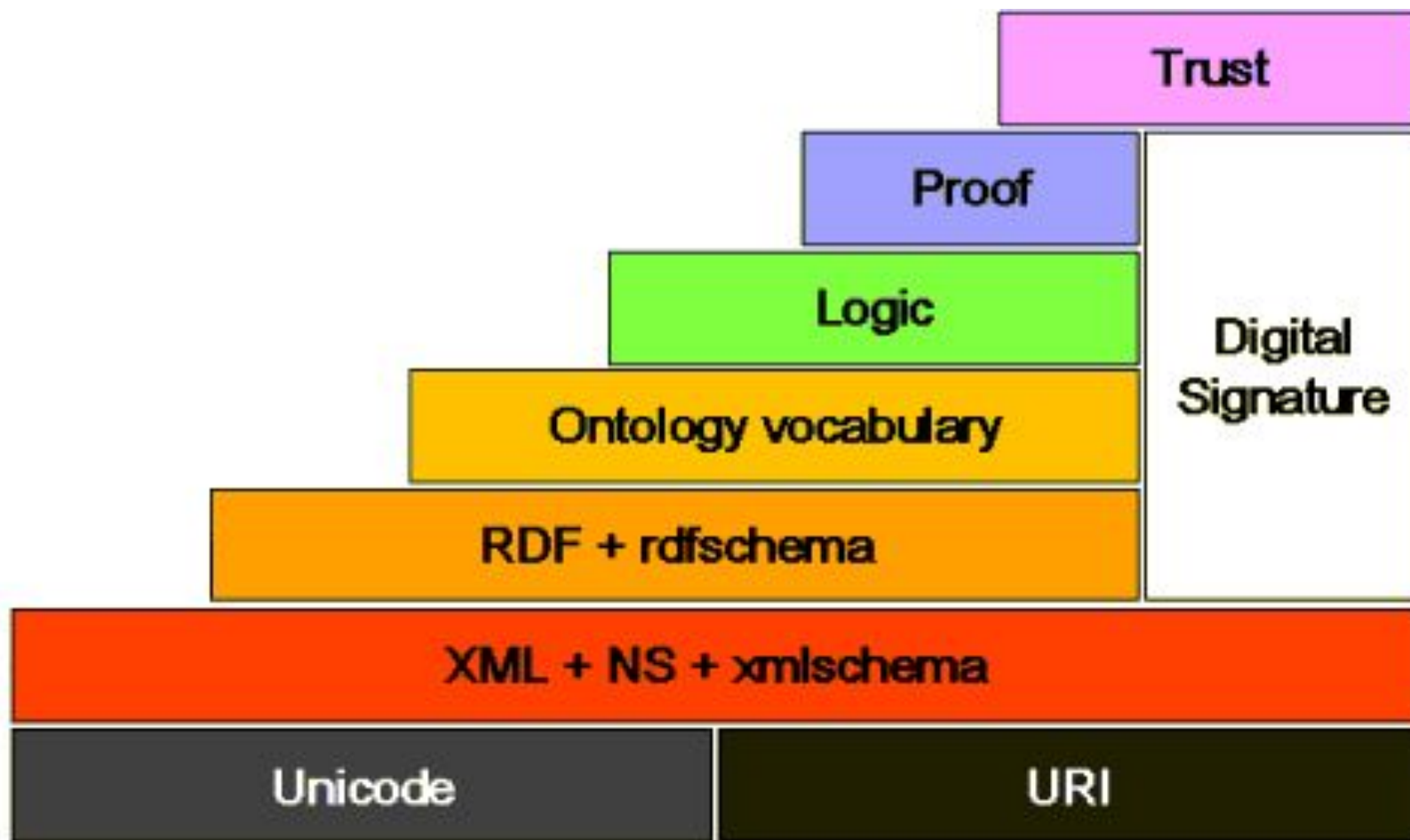
Các công nghệ Web ngữ nghĩa

- Tập hợp các quy chuẩn cùng với các triển khai, các bộ dữ liệu liên kết;
- Hỗ trợ chia sẻ dữ liệu và tri thức trong môi trường Web.
- Các công nghệ cơ bản sử dụng Lô-gic tính toán làm nền tảng ngữ nghĩa
 - Đơn giản, hành vi xác định, có thể có giải thuật suy diễn hiệu quả.
 - - *Không có cơ chế xác suất*
- Sử dụng đồ thị để biểu diễn tri thức
 - Đơn giản, công cụ hỗ trợ tốt
 - - *Có thể quá đơn giản*

Các công nghệ Web ngữ nghĩa₍₂₎

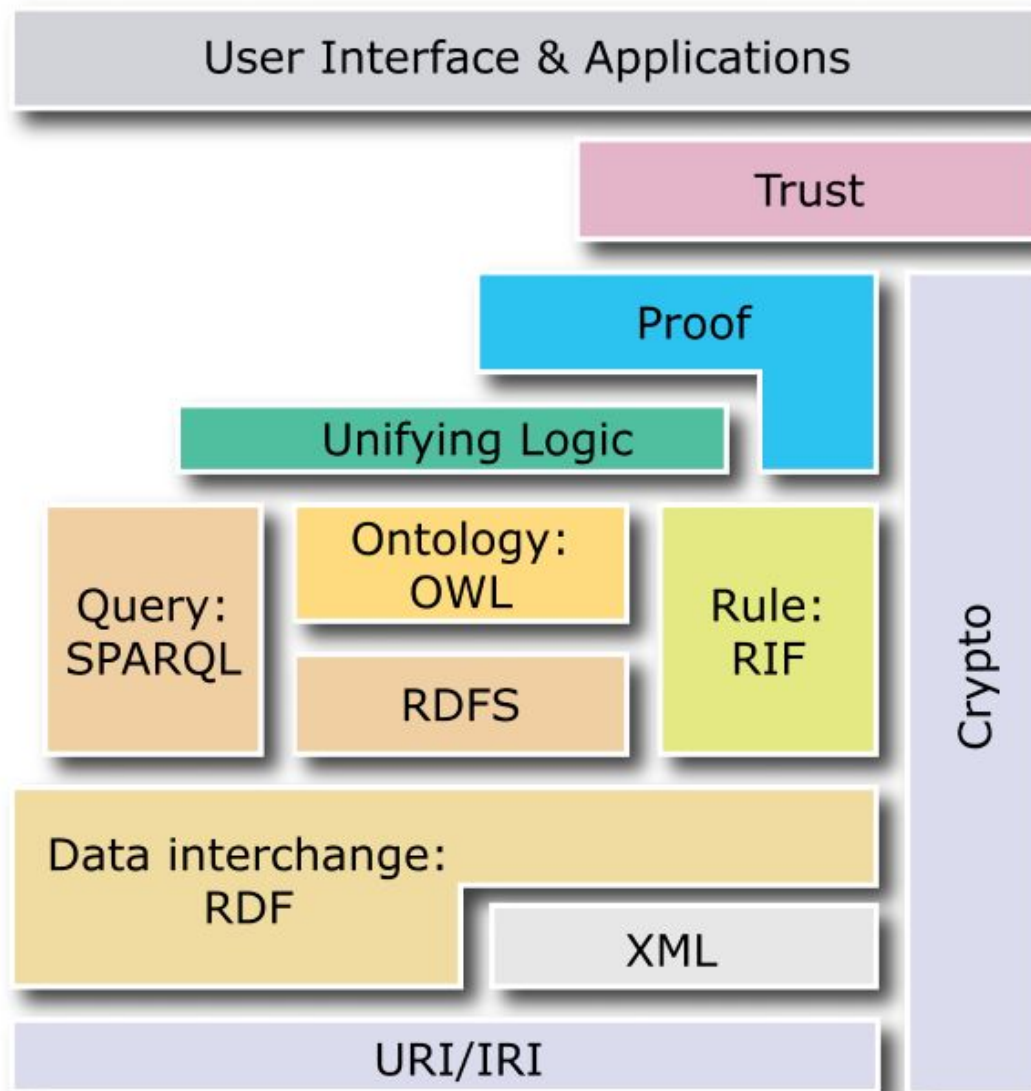
- Các quy chuẩn W3C
 - RDF, RDFS, RDFa, OWL, SPARQL, RIF, SHACL, v.v..
- Các công cụ và hệ thống ứng dụng - Thương mại, miễn phí và mã nguồn mở
 - Trình soạn thảo ontology, lưu trữ bộ-3, mô tơ suy diễn, v.v.
 - Ví dụ, Apache Jena Fuseki
- Các ontologies và các bộ dữ liệu
 - Foaf, Dublin Core, DBpedia, SKOS, PROV, v.v.
- Các hạ tầng dịch vụ
 - Tìm kiếm, dịch vụ liên kết, v.v..
- Các quy chuẩn bên ngoài W3C: Schema.org, Freebase, Open Graph, ...

Ngăn xếp công nghệ Web ngữ nghĩa



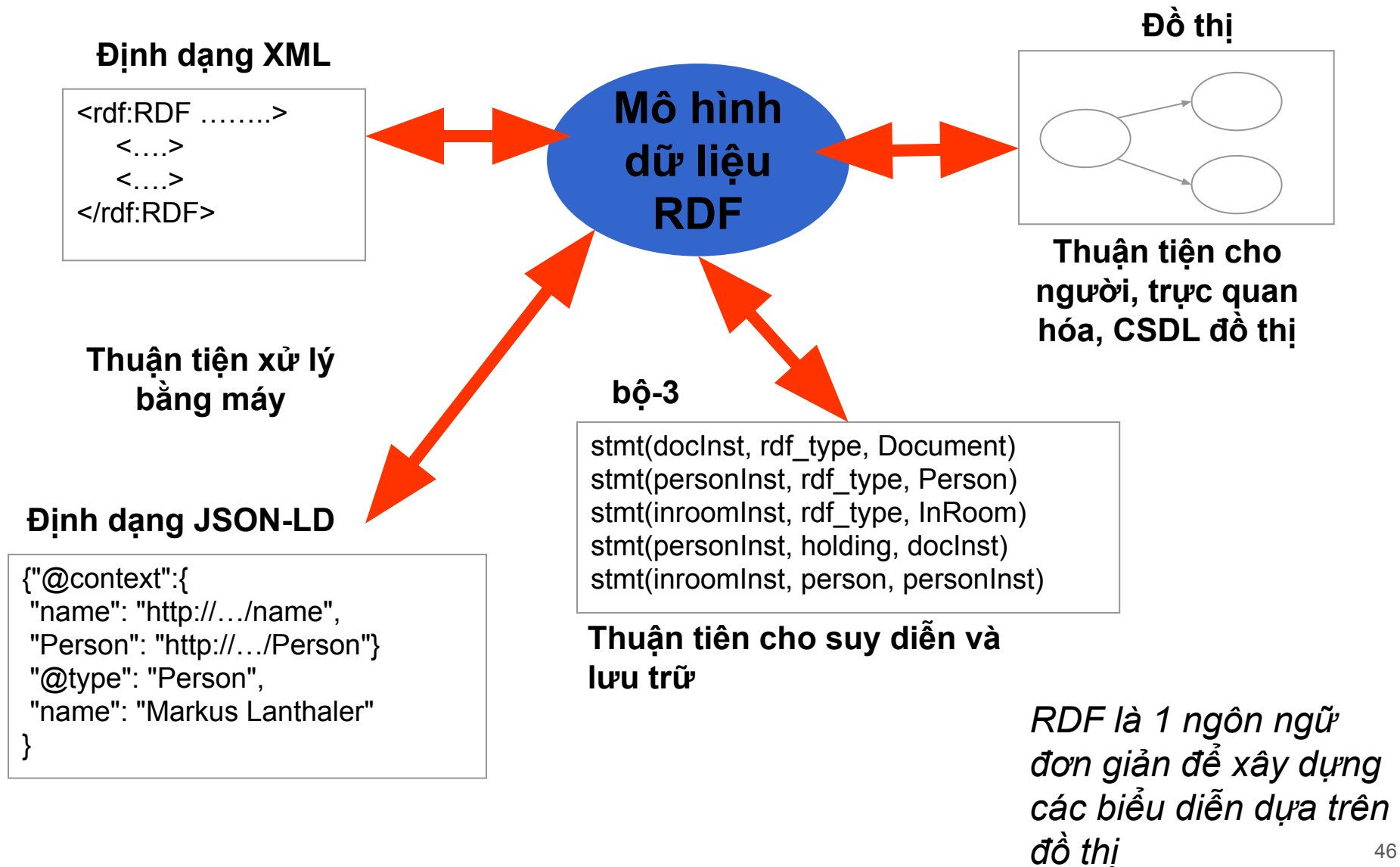
[Tim Berners-Lee]

Ngăn xếp công nghệ Web ngữ nghĩa₍₂₎



[W3C]

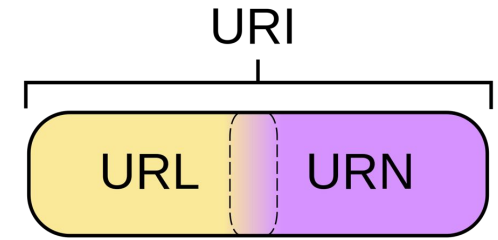
RDF là ngôn ngữ Web ngữ nghĩa đầu tiên



Mô hình dữ liệu RDF

- Tài liệu RDF là tập không thứ tự các câu, mỗi câu có cú pháp đơn giản bao gồm chủ ngữ (subject), thuộc tính (predicate) và giá trị (object)
 - Câu RDF = Bộ-3
- Mỗi bộ-3 có thể được biểu diễn như 1 cạnh nối 2 đỉnh có nhãn trong đồ thị
- Các câu mô tả các thuộc tính của các tài nguyên Web
- Tài nguyên là thực thể được xác định bằng URI
 - - Bài viết, hình ảnh, đoạn văn trên Web
 - Ví dụ, https://vi.wikipedia.org/wiki/Alan_Turing
 - hoặc 1 cuốn sách: isbn://5031-4444-3333
- Các thuộc tính đồng thời cũng là các tài nguyên (URIs)

URIs giữ vai trò nền tảng



- URI = (Uniform Resource Identifier)
 - Chuỗi định danh thống nhất.
 - Bao gồm tập hợp các tên/địa chỉ là những chuỗi ký tự ngắn tham chiếu các tài nguyên
 - và URLs (Uniform Resource Locators, được sử dụng cho các tài nguyên Web).
- URIs nhìn giống URLs, thường có chỉ số đoạn để chỉ tới 1 phần của tài liệu
 - `http://foo.com/bar/mumble.html#pitch`

URIs giữ vai trò nền tảng₍₂₎

- URIs có ý nghĩa xác định, không nhập nhằng về nghĩa như các từ ngôn ngữ tự nhiên
 - Web cung cấp 1 không gian tên toàn cầu.
- Chúng ta có thể sử dụng URI để tham chiếu bất cứ thứ gì, ví dụ, 1 khái niệm, thực thể, sự kiện hoặc quan hệ.
- Chúng ta thường giả định 2 thứ có URIs giống nhau (có thể xuất hiện ở hai ngữ cảnh khác nhau) là một.

Nghĩa của URI là gì?

- Trong nhiều trường hợp URIs chỉ định 1 tài nguyên Web
- Đôi khi là những thực thể trong thế giới thực
 - Ví dụ, <https://www.hust.edu.vn/> là 1 Đại học hàng đầu của Việt Nam.
- Các URIs tốt thường không thay đổi
 - <http://www.w3.org/Provider/Style/URI>

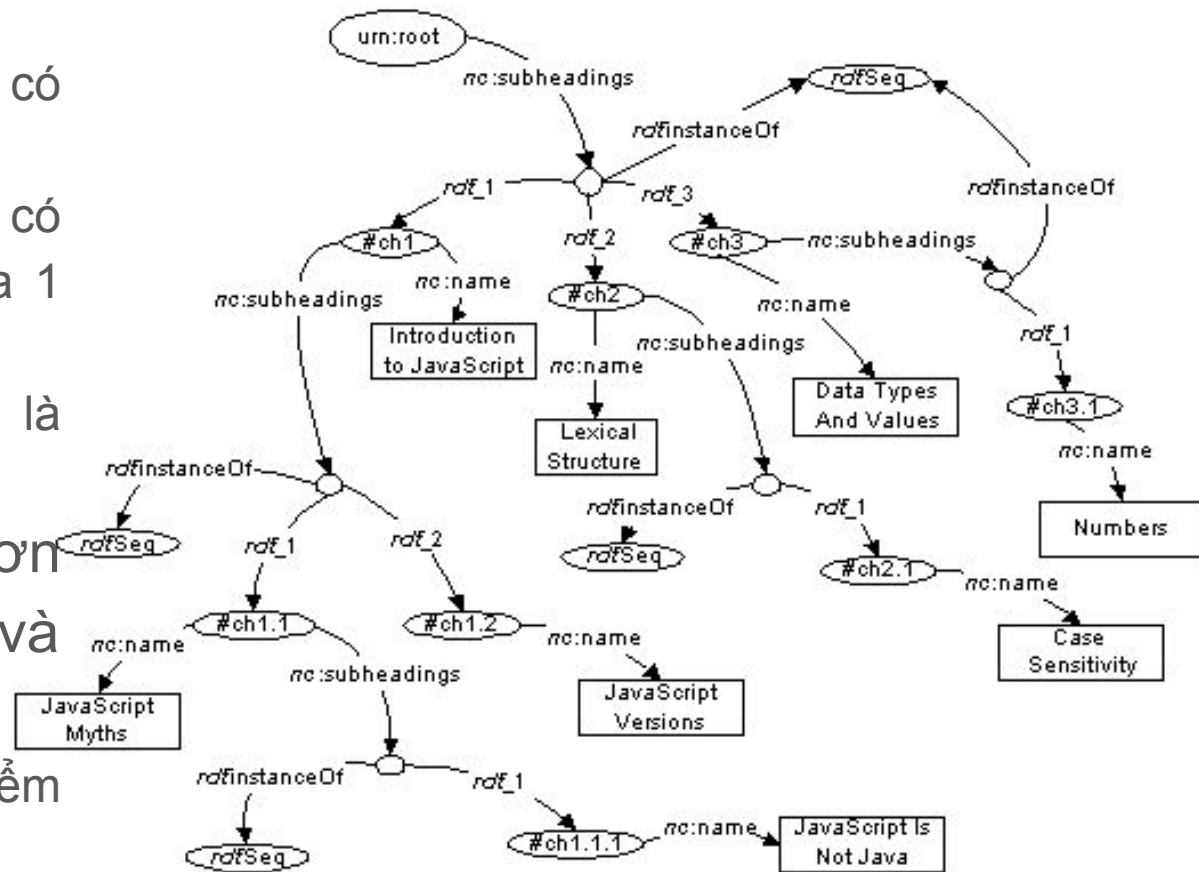
Đồ thị RDF

- Biểu diễn chủ thể và đối tượng như các nút, khẳng định như cạnh có hướng

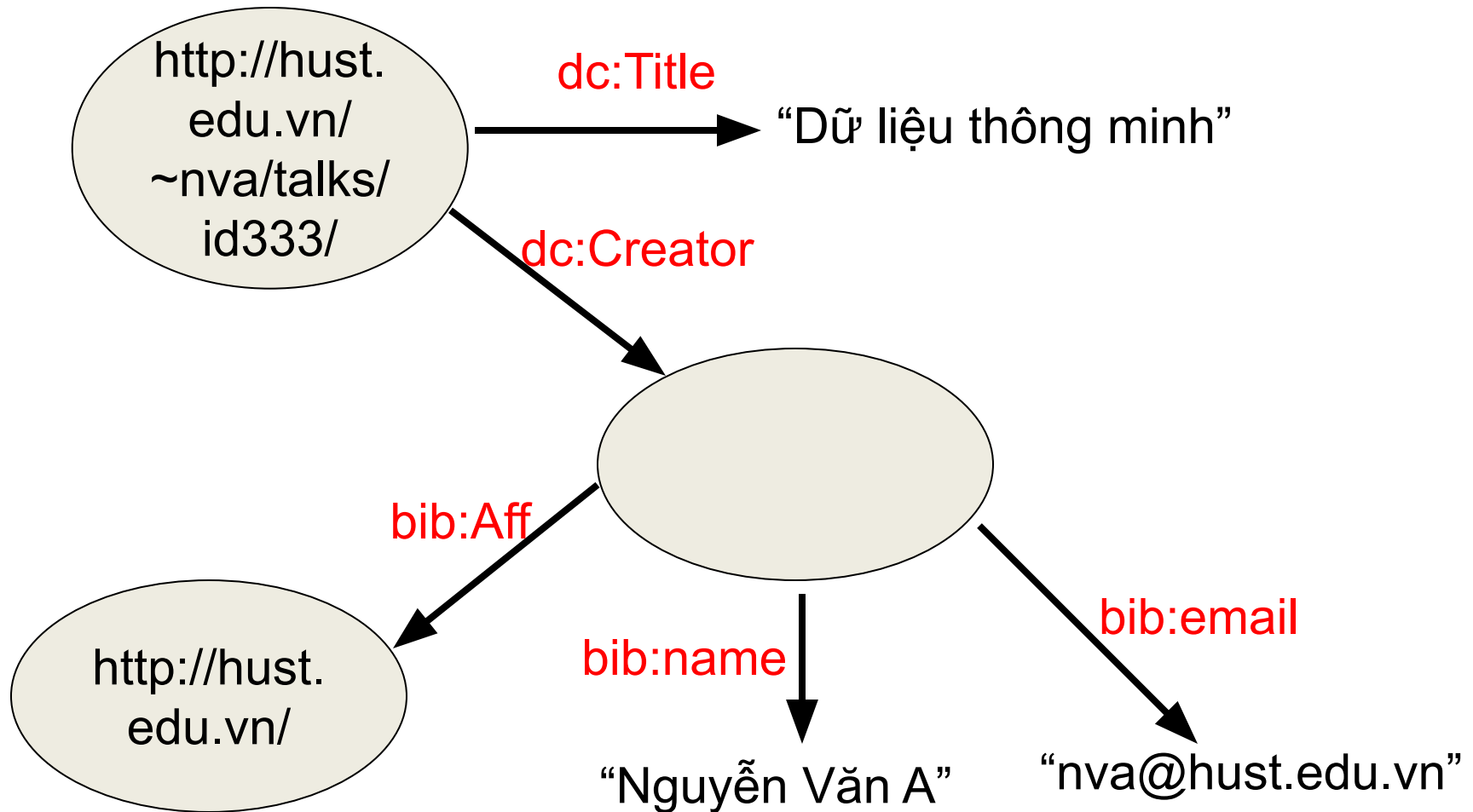
- Thu được đồ thị có hướng có gán nhãn
- Chủ thể của 1 bộ-3 có thể là đối tượng của 1 bộ-3 khác
- Đối tượng có thể là chuỗi ký tự

- Đồ thị đơn giản hơn CSDL quan hệ và CSDL đối tượng

- Hàm chứa cả ưu điểm và nhược điểm



Ví dụ 1.3. Đồ thị RDF đơn giản



Các dữ kiện trong ví dụ chỉ có tính minh họa

Lưu trữ

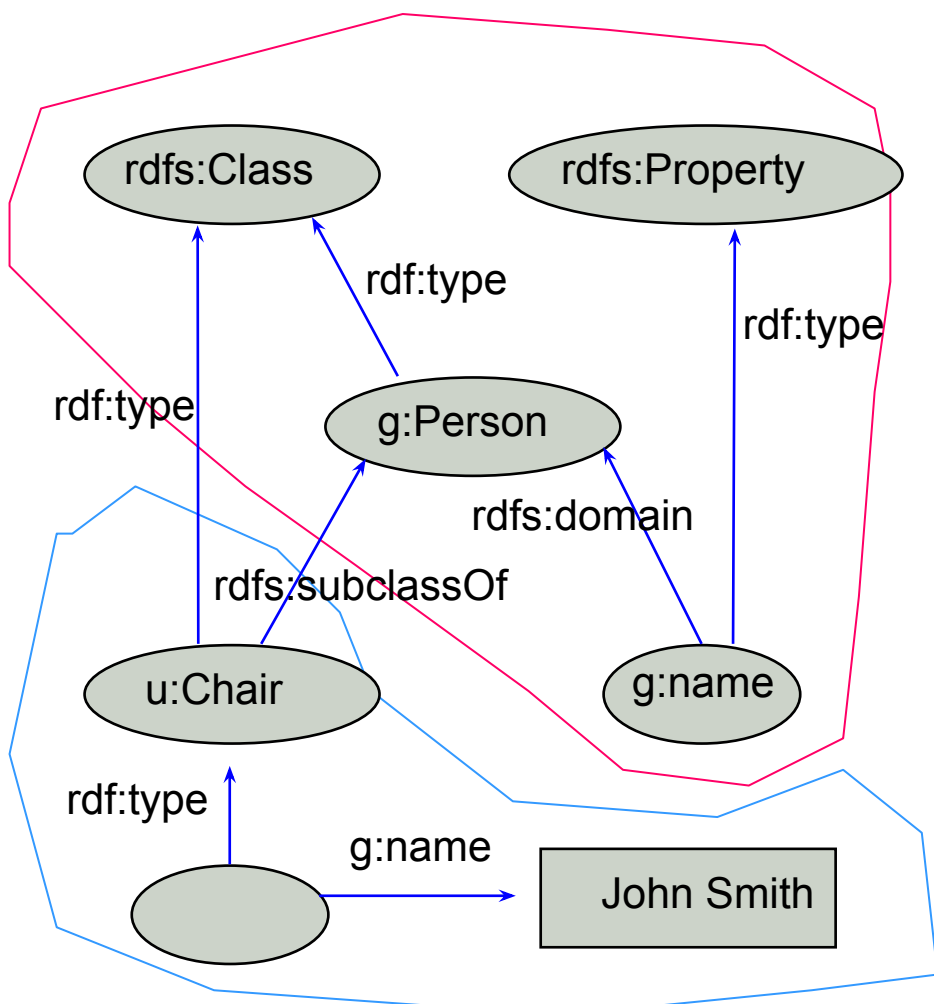
- Đồ thị là 1 mô hình trừu tượng, chúng ta cần lưu trữ nó như văn bản để xử lý, ví dụ, soạn thảo, trao đổi, phân tích, v.v...
- Có nhiều quy chuẩn lưu trữ cho dữ liệu RDF
- Các quy chuẩn quan trọng nhất: XML, Turtle, N-Triples, JSON-LD
 - Hầu hết các công cụ Web ngữ nghĩa đều có thể đọc và viết với các định dạng này.

Chúng ta sẽ chủ yếu thực hành với định dạng Turtle

Lược đồ RDF (RDFS)

- Lược đồ RDF bổ xung các cây phân cấp cho các lớp & các thuộc tính
 - Lớp con (subClass) và thuộc tính con (subProperty).
- Thêm các siêu dữ liệu
 - Ví dụ, ràng buộc miền và khoảng trên các thuộc tính.
- Nhiều hệ quản trị CSTT có thể nhập và xuất tài liệu RDFS.

RDF và lược đồ RDF



Thông tin mức Lược đồ

```
@prefix rdf:
  http://www.w3.org/1999/02/22-rdf-syntax-ns# .
@prefix rdfs:
  http://www.w3.org/2000/01/rdf-schema# .
@prefix g: http://schema.org/gen .
@prefix u: http://schema.org/univ .

g:name rdf:type rdfs:Property;
  rdfs:domain g:Person .

u:Chair rdfs:subClassOf g:Person .
```

Thông tin mức dữ kiện

```
_john rdf:type u:Chair;
  g:name "John Smith" .
```

*!Có thể mô tả bất
cứ điều gì, kể cả
nội dung sai.*

RDFS hỗ trợ các suy diễn đơn giản

- Một ontology RDFS và một vài câu RDF có thể ngầm định thêm các câu RDF khác
 - Được coi như 1 phần của mô hình dữ liệu
- *(Không đúng với dữ liệu XML)*

```
@prefix rdfs: <http://www...>.
@prefix : <...genesis.n3>.
:parent rdfs:domain :Person;
        rdfs:range :Person.
:mother
  rdfs:subProperty parent;
  rdfs:domain :Woman.
:eve :mother :cain.
```



```
:parent a rdf:Property.
:Person a rdf:Class.
:Woman rdfs:subClassOf Person.
:mother a rdf:Property.
:eve a :Person;
      a :Woman;
      :parent :cain.
:cain a :Person.
```


Liệu RDF(S) có tốt hơn XML?

- Phụ thuộc vào từng tình huống ứng dụng cụ thể
- Mô hình XML
 - Cây phân cấp chặt chẽ;
 - Các ứng dụng có thể dựa trên vị trí trong cây;
 - Cú pháp và cấu trúc tương đối đơn giản;
 - Tương đối khó kết hợp.
- Mô hình RDF
 - Tập hợp linh động của các liên kết;
 - Các ứng dụng có thể thực hiện các tìm kiếm tương tự CSDL;
 - Tương đối khó khôi phục cấu trúc;
 - Dễ dàng hợp nhất các mô hình thành 1 mô hình lớn;
 - Tốt cho việc tích hợp thông tin từ nhiều nguồn.

Một số hạn chế của RDFS

- Không có khả năng mô tả chi tiết các tài nguyên
 - Không thể giới hạn ràng buộc domain và range trong phạm vi hẹp hơn: Không thể mô tả range của hasChild là cat khi áp dụng cho cat và là dog khi áp dụng cho dog.
 - Không có ràng buộc tồn tại/cơ số: Không thể mô tả 1 cái xe đạp có 2 bánh.
 - Không có các thuộc tính bắc cầu, nghịch đảo hoặc đối xứng: Không thể nói isPartOf là 1 thuộc tính bắc cầu, hasPart là nghịch đảo của isPartOf hoặc touches là 1 thuộc tính đối xứng.
- => Cần các từ khóa mới để mở rộng khả năng mô tả

Ngôn ngữ OWL

- Dự án DARPA, DAML+OIL, tiền thân của OWL
- OWL được công bố như 1 quy chuẩn W3C, 2/10/04
- Ba phiên bản OWL đã được xác định theo thứ tự giảm dần độ phức tạp cùng với khả năng diễn đạt
 - OWL Full - Đầy đủ nhất
 - OWL DL - (Lô-gic mô tả) giới thiệu các giới hạn
 - OWL Lite - Ngôn ngữ mức bắt đầu, hướng tới tính đơn giản và dễ triển khai
- OWL 2 trở thành quy chuẩn W3C trong năm 2009, được cập nhật năm 2012

OWL \Leftrightarrow RDF

- Tài liệu OWL là tập các câu RDF
 - OWL xác định ngữ nghĩa cho các câu cụ thể.
- Thêm các tính năng thông dụng cho Lô-gic mô tả, ví dụ, ràng buộc cơ sở, định nghĩa lớp, tương đương, các lớp không giao, v.v..
- Hỗ trợ quản lý ontologies như những đối tượng (ví dụ, import, phiên bản, ...).
- Suy diễn OWL phức tạp hơn nhiều so với suy diễn RDFS

OWL \Leftrightarrow RDF₍₂₎

- RDF cho phép mô tả dữ liệu.
- RDFS thêm vào khả năng mô tả dữ liệu ở mức lược đồ.
- OWL cho phép nhiều thông tin mức lược đồ hơn.
- Chúng ta thường sử dụng RDFS và OWL để định nghĩa ontologies lĩnh vực (các lược đồ).
- Và sau đó sử dụng những ontologies đó để đưa ra thông tin về các đối tượng.

Dữ liệu ngữ nghĩa nhúng trong HTML

- Nhúng dữ liệu ngữ nghĩa trong HTML cho phép cả người và máy tính cùng hiểu văn bản
 - RDFa là 1 quy chuẩn nhúng RDF trong HTML như các thuộc tính của thẻ.
 - JSON-LD là quy chuẩn để nhúng RDF trong 1 định dạng tương thích với JSON.
- Facebook tìm kiếm các câu RDFa được nhúng sử dụng bộ từ vựng đồ thị mở (open graph - og).

Các công cụ phần mềm

- Có nhiều trình soạn thảo Ontology
 - Giao diện trực quan, hỗ trợ suy diễn
 - Ví dụ, Topbraid Composer, Protégé
 - Tuy nhiên có thể biên soạn các bộ-3 bằng các trình soạn thảo văn bản.
- Lưu trữ bộ-3: Cơ sở dữ liệu bộ-3
 - Thường hỗ trợ các API riêng và thường có cổng tiếp nhận truy vấn SPARQL
 - Có thể thực hiện suy diễn: Tức thì hoặc theo yêu cầu
 - Ví dụ, Apache Jena Fuseki, GraphDB
- Nền tảng và thư viện: Cho các ngôn ngữ lập trình
 - Java: Jena, hỗ trợ lưu trữ, SPARQL, suy diễn, v.v.

Nội dung

1.1. Khái niệm Web ngữ nghĩa

1.2. Web ngữ nghĩa hiện nay

1.3. Đồ thị tri thức

1.4. Tổng quan công nghệ Web ngữ nghĩa

1.5. Ontology và một số khái niệm quan trọng



Khái niệm Ontology trong triết học

- Từ Ontology trong triết học được sử dụng lần đầu tiên trong thế kỷ 17.
 - Chỉ tồn tại ở dạng số ít;
 - Không có ontologies trong triết học.
- Được định nghĩa bởi các nhà triết học như bản chất tự nhiên của sự sống và sự tồn tại
 - *(Bản thể luận)*
 - Từ thế kỷ 5 TCN, Empedocles, học thuyết về sự hình thành của thế giới từ 4 nguyên tố - Đất, lửa, nước và không khí
 - Được sử dụng như công cụ để phân loại và hệ thống hóa tri thức

Trong khoa học máy tính, có nhiều cách định nghĩa khái niệm ontology

Một số định nghĩa ontology trong KHMT

- Ontology là 1 tập từ khóa có cấu trúc phân cấp để mô tả 1 lĩnh vực và có thể được sử dụng như cấu trúc khung cho 1 cơ sở tri thức
 - *An ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge base. (Swartout, Patil, Knight, Russ)*
- Trong môi trường Web ngữ nghĩa - Bộ từ vựng xác định các khái niệm và các mối quan hệ được sử dụng để mô tả 1 lĩnh vực. [W3.org]
 - Ontology \approx Bộ từ vựng (Vocabulary)
 - Từ ontology thường được sử dụng cho các bộ từ vựng phức tạp và có tính hình thức, còn khái niệm bộ từ vựng thường được sử dụng khi không yêu cầu các ràng buộc hình thức.

Một số định nghĩa ontology trong KHMT₍₂₎

Gruber (1993)

"Ontology là 1

đặc tả hình thức \Rightarrow biểu diễn để xử lý bằng máy tính

của 1 **hệ khái niệm** \Rightarrow các khái niệm và các quan hệ

dùng chung \Rightarrow dựa trên các thỏa thuận

của 1 **lĩnh vực** ứng dụng \Rightarrow ngữ cảnh của khái niệm

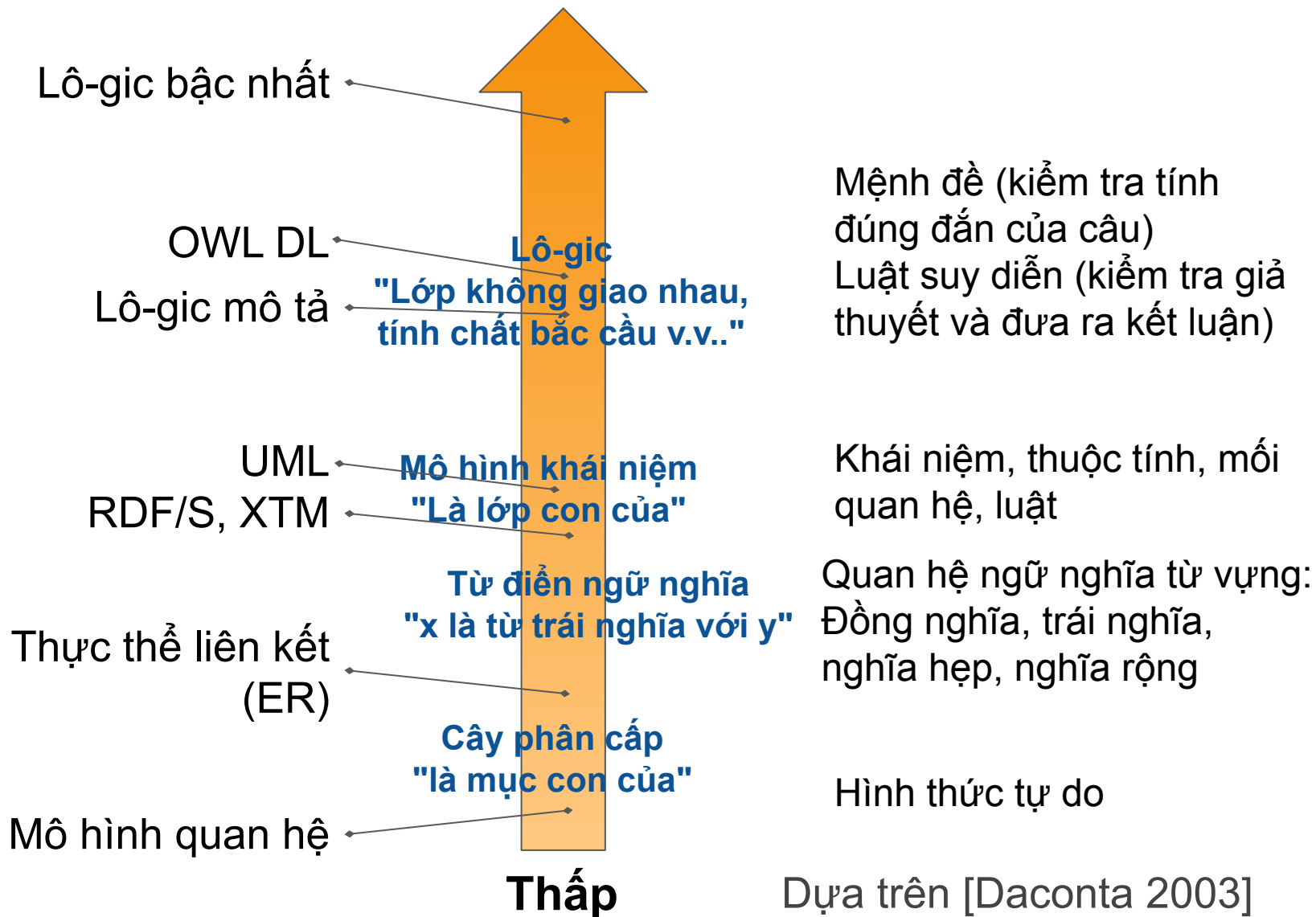
Các mức ontology

- Cây phân cấp - Taxonomy (tri thức với cấu trúc cây tối thiểu), quan hệ là mục con của.
- Từ điển ngữ nghĩa - Thesaurus, sử dụng tập quan hệ hữu hạn (đồng nghĩa, trái nghĩa, mở rộng, thu hẹp, v.v..), ví dụ WordNet.
- Mô hình khái niệm - Conceptual Model, biểu diễn các tri thức đơn giản.
- Lý thuyết Lô-gic - Các tri thức phức tạp.

Các mức ontology⁽²⁾

Hàm lượng ngữ nghĩa

Cao



Dựa trên [Daconta 2003]

So sánh sơ đồ quan hệ và ontology

- Mục đích chính của sơ đồ quan hệ là tổ chức dữ liệu trong cơ sở dữ liệu
 - Các mối quan hệ giữa các thực thể được ngầm định bởi đối tượng sử dụng: người/chương trình máy tính.
 - Trong CSDL quan hệ chỉ có liên kết khóa chính-khóa ngoại
 - Nếu người/chương trình máy tính không hiểu ý nghĩa của dữ liệu thì không biểu diễn được các quan hệ.
- Ontology
 - Các mối quan hệ được định nghĩa tường minh bằng các ngôn ngữ hình thức để cả người và chương trình máy tính đều có thể biểu diễn được

Các khía cạnh ngôn ngữ

Mỗi ngôn ngữ thường bao gồm 3 thành phần:

- 1. Cú pháp:** Syntax - các quy ước ngữ pháp để tạo các câu trong ngôn ngữ.
- 2. Ngữ nghĩa:** Semantic - ánh xạ các câu tới ý nghĩa (có thể là giá trị chân lý theo lý thuyết).
- 3. Ngữ dụng:** Pragmatics - phần còn lại, các khía cạnh ứng dụng (cách vận dụng ngôn ngữ, hiểu biết về thế giới, v.v..).

Ngôn ngữ hình thức: Cú pháp

- Các URIs (*Uniform Resource Identifier*) chỉ các lớp, các thuộc tính, các đối tượng, ví dụ:

http://live.dbpedia.org/resource/Alan_Turing

<http://schema.org/Scientist>

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

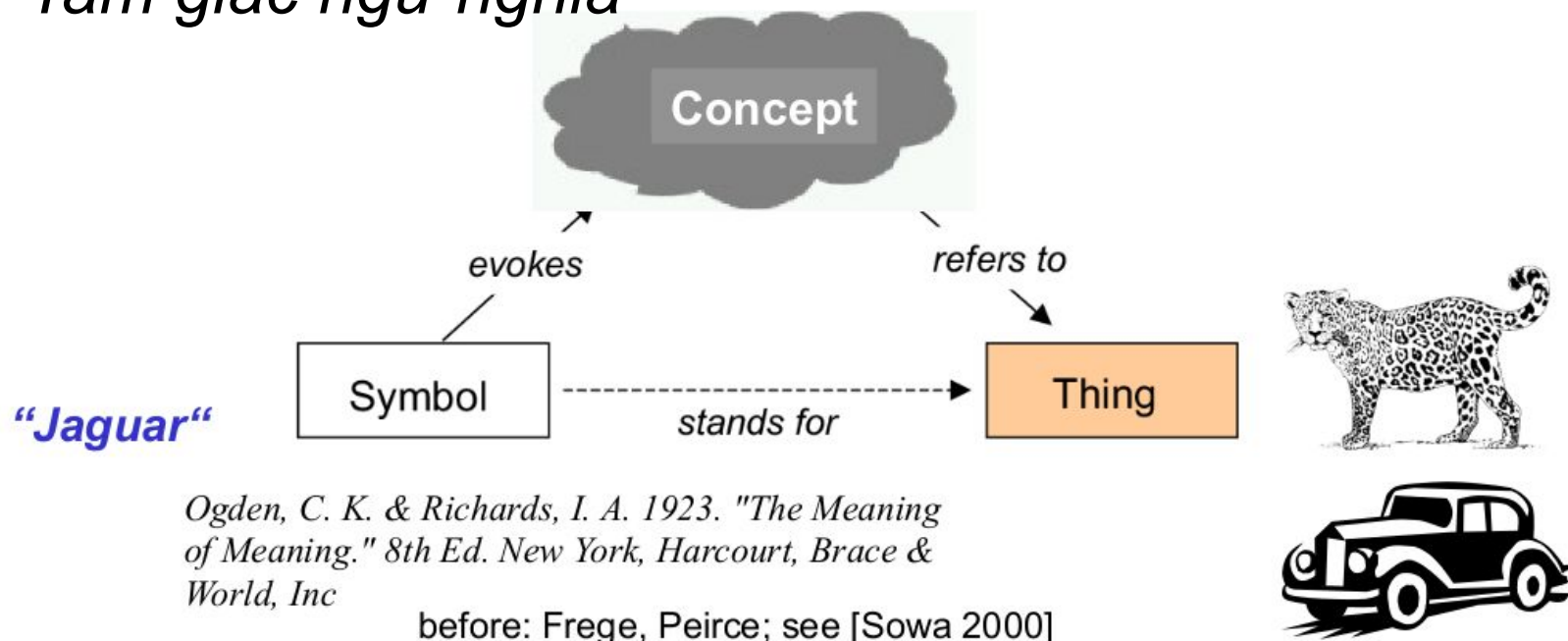
- Sử dụng chuỗi ký tự cho nội dung văn bản
- Sử dụng bộ-3 để tạo các câu, ví dụ:

dbpedia:Alan_Turing rdfs:type schema:Scientist .

- “Alan Turing is a scientist” - "Alan Turing là 1 nhà khoa học"

Ngôn ngữ hình thức: Ngữ nghĩa

Tam giác ngữ nghĩa

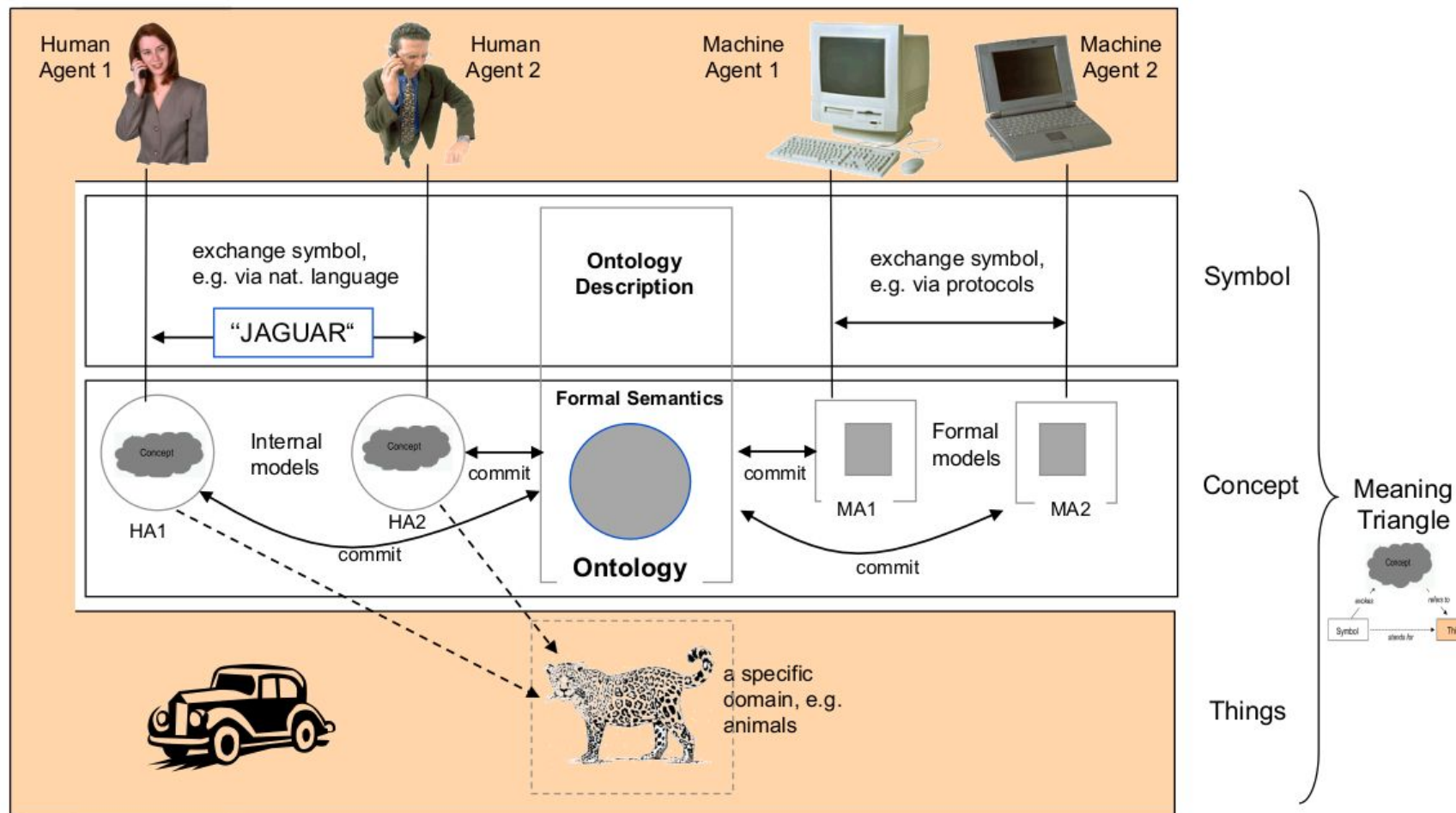


[Carole Goble, Nigel Shadbolt, Ontologies and the Grid Tutorial]

Khi người tiếp nhận thông tin, các khái niệm là trung gian kết nối từ (hoặc tối thiểu là ký hiệu) với các thực thể.

Ngôn ngữ hình thức: Ngữ nghĩa (2)

Khi máy tiếp nhận thông tin



Maedche et al., 2002

Máy tính xử lý dựa trên các biểu diễn hình thức

Ngôn ngữ hình thức: Ngữ nghĩa₍₃₎

Ngữ nghĩa trong Web ngữ nghĩa

- Ánh xạ URI tới các biểu diễn của thực thể mà nó chỉ định trong "thế giới" (Thực và ảo) trong 1 nền tảng hình thức
 - Các URI là duy nhất
- Máy tính có thể xử lý ngữ nghĩa trên nền tảng hình thức, điển hình là lô-gic.
- Cho phép thực hiện suy diễn, ví dụ:
 - Quan hệ parent (là phụ mẫu) là nghịch đảo của quan hệ child (là con);
 - `schema:parent owl:inverse schema:child` .

Ngôn ngữ hình thức: Ngữ dụng

- Các giao thức, ngữ cảnh, v.v..
 - Một số vấn đề được xử lý bởi Giao thức Web (GET, POST);
 - Một số được xử lý bằng các giao thức chuyên dụng (ví dụ, truy vấn SPARQL).
- Các ứng dụng
 - Ví dụ, các cơ sở tri thức bách khoa toàn thư (ví dụ DBpedia) để hỗ trợ xác định các thực thể phổ biến.

