Lecture #A

NEWM N510: Web-Database Concepts

Semantic Web

kharrazi@iupui.edu
http://www.info510.com

Lecture in a Nutshell

1. Definition
2. Purpose
3. Relationship to The Hypertext Web
4. Skeptical Reactions
5. Components
6. Challenges
7. Projects
8. Example

1. Definition

- Semantic Web is a group of methods and technologies to allow machines to understand the meaning *(semantics)* of information on the World Wide Web

- According to the original vision, the availability of machine-readable metadata would enable automated agents and other software to access the Web more intelligently. The agents would be able to perform tasks automatically and locate related information on behalf of the user.

- While the term "Semantic Web" is not formally defined it is mainly used to describe the model and technologies proposed by the W3C. These technologies include the Resource Description Framework (RDF), a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS) and the Web Ontology Language (OWL), all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain.

- Many of the technologies proposed by the W3C already exist and are used in various projects. The Semantic Web as a global vision, however, has remained largely unrealized and its critics have questioned the feasibility of the approach.

2. Purpose

- Humans are capable of using the Web to carry out tasks such as finding the French word for "directory," reserving a library book, and searching for a low price for a DVD. However, one computer cannot accomplish all of these tasks without human direction, because web pages are designed to be read by people, not machines.

- The semantic web is a vision of information that is understandable by computers, so computers can perform more of the tedious work involved in finding, combining, and acting upon information on the web.

- Tim Berners-Lee originally expressed the vision of the semantic web as:

  "I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize."

Purpose cont.

- Semantic Web application areas are experiencing intensified interest due to the rapid growth in the use of the Web, together with the innovation and renovation of information content technologies.

- The Semantic Web is regarded as an integrator across different content and information applications and systems, and provide mechanisms for the realization of Enterprise Information Systems.

- Often the terms 'Semantics', 'metadata', 'ontologies' and 'Semantic Web' are used inconsistently. In particular, these terms are used as everyday terminology by researchers and practitioners, spanning a vast landscape of different fields, technologies, concepts and application areas. Furthermore, there is confusion with regards to the current status of the enabling technologies envisioned to realize the Semantic Web.

- Tim Berners-Lee has described semantic web part of 'Web 3.0':

  "People keep asking what Web 3.0 is. I think maybe when you've got an overlay of scalable vector graphics - everything rippling and folding and looking misty — on Web 2.0 and access to a semantic Web integrated across a huge space of data, you'll have access to an unbelievable data resource..."

3. Relationship to The Hypertext Web

❑ Limitations of HTML

- Many files on a typical computer can be loosely divided into: Documents like mail messages, reports, and brochures are read by humans. Data, like calendars, address-books, playlists, and spreadsheets are presented using an application program which lets them be viewed, searched and combined in many ways.

- Currently, the World Wide Web is based mainly on documents written in Hypertext Markup Language (HTML), a markup convention that is used for coding a body of text interspersed with multimedia objects such as images and interactive forms.

- Metadata tags provide a method by which computers can categorize the content of web pages:

```
<meta name="keywords" content="Hospital Information System, HIS, Electronic Health Record, EHR, Electronic Medical Record">
<meta name="description" content="New Commercial EHR Systems">
<meta name="author" content="John Doe – HIMMS">
```

Relationship to The Hypertext Web cont.

- With HTML and a tool to render it (perhaps web browser or a user agent), one can create and present a page that lists admitted patients.

  The HTML of this list page can make simple, document-level assertions such as "this document's title is ‘Patient List’", but there is no capability within the HTML itself to assert unambiguously that, for example, patient HK783545e is a Male with a chief complaint of Abdominal Pain, or that he is 45 yrs old.

  Rather, HTML can only say that the span of text "HK783545e" is something that should be positioned near “Male” and “Abdominal Pain”.

  There is no way to say “This is a demographic list” or even to establish that “Male” is a kind of gender or that “Abdominal Pain” is a chief complaint. There is also no way to express that these pieces of information are bound together in describing a patient, distinct from other patients perhaps listed on the page.

- Microformats represent unofficial attempts to extend HTML syntax to create machine-readable semantic markup about objects such formatted name, telephone numbers and etc.

Relationship to The Hypertext Web cont.

❑ Semantic Web Solutions

- The Semantic Web takes the solution further. It involves publishing in languages specifically designed for data: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML describes documents and the links between them. RDF, OWL, and XML, by contrast, can describe arbitrary things such as people, meetings, or airplane parts → resulting network of Linked Data the Giant Global Graph, in contrast to the HTML-based World Wide Web.

- These technologies are combined in order to provide descriptions that supplement or replace the content of Web documents. Thus, content may manifest itself as descriptive data stored in Web-accessible databases, or as markup within documents (particularly XHTML interspersed with XML, or purely in XML with layout or rendering XSL stored separately).

- The machine-readable descriptions enable content managers to add meaning to the content → a machine can process knowledge itself:

```
<item>appendicitis</item>
<item rdf:about="http://en.wikipedia.org/wiki/Appendicitis">appendicitis</item>
```

4. Skeptical Reactions

❑ Practical Feasibility

- Critics (e.g. Which Semantic Web?) question the basic feasibility of a complete or even partial fulfillment of the semantic web.

  Metacrap → human behavior and personal preferences. For example, people lie: they may include spurious metadata into Web pages in an attempt to mislead Semantic Web engines that naively assume the metadata's veracity. This phenomenon was well-known with metatags that fooled the search engine ranking algorithms into elevating the ranking of certain Web pages.

❑ Censorship and Privacy

- Text-analyzing techniques can now be easily bypassed by using other words, metaphors for instance, or by using images in place of words. An advanced implementation of the semantic web would make it much easier for governments to control the viewing and creation of online information, as this information would be much easier for an automated content-blocking machine to understand → geo location meta-data: very little anonymity associated with the authorship of articles.

Skeptical Reactions cont.

❑ Doubling Output Formats

- It would be much more time-consuming to create and publish content because there would need to be two formats for one piece of data: one for human viewing and one for machines. However, many web applications in development are addressing this issue by creating a machine-readable format upon the publishing of data or the request of a machine for such data. The development of microformats has been one reaction to this kind of criticism.

❑ Need

- The idea of a semantic web, able to describe, and associate meaning with data, necessarily involves more than simple XHTML mark-up code. It is based on an assumption that, in order for it to be possible to endow machines with an ability to accurately interpret web content, far more than the mere ordered relationships involving letters and words is necessary as underlying infrastructure → Otherwise, most of the supportive functionality would have been available in Web 2.0, and it would have been possible to derive a semantically capable Web with minor additions.

5. Components

- The semantic web comprises the standards and tools of XML, XML Schema, RDF, RDF Schema and OWL that are organized in the Semantic Web Stack. The OWL Web Ontology Language Overview describes the function and relationship of each of these components of the semantic web:

  XML provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within.

  XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.

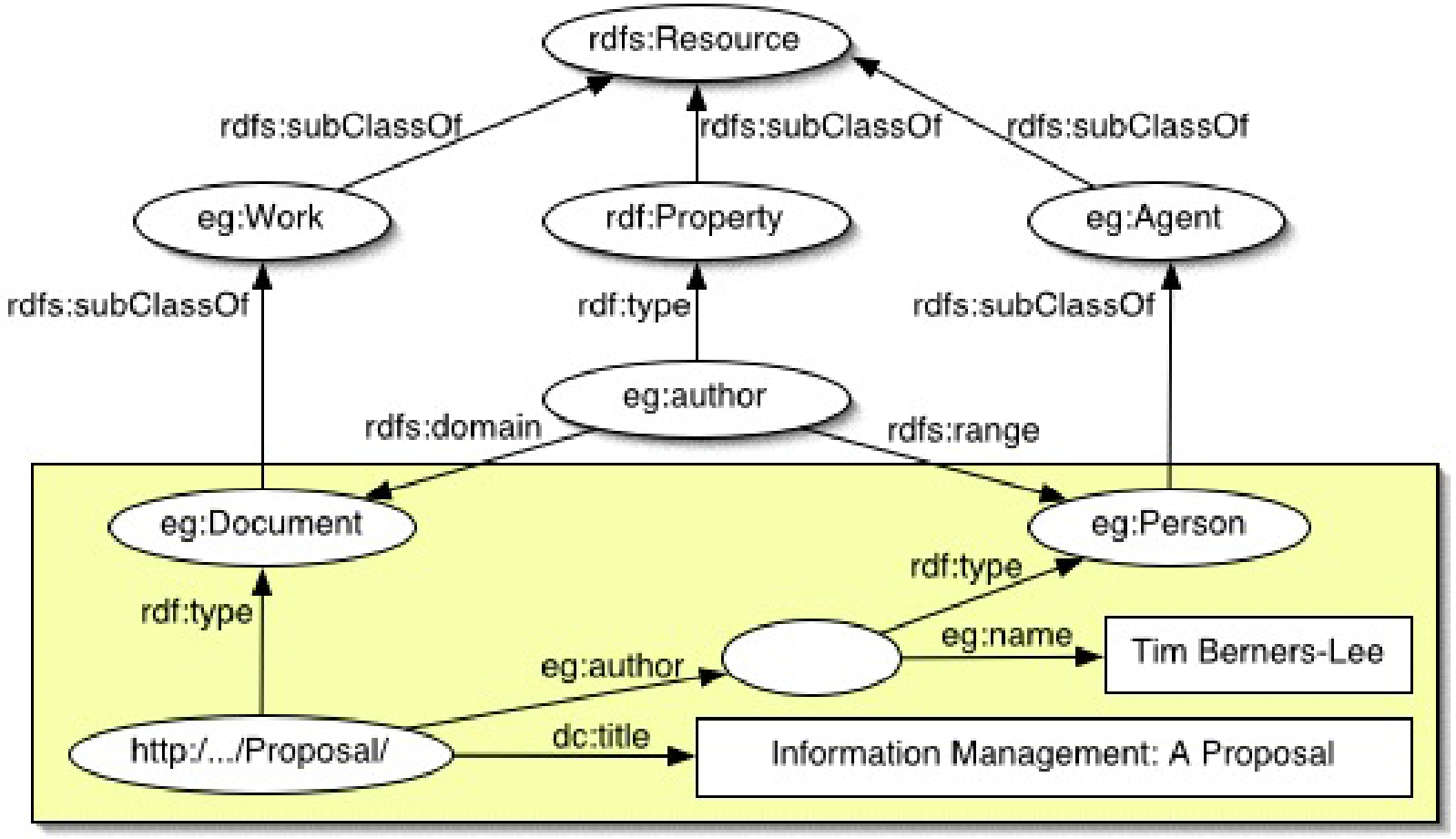
  RDF is a simple language for expressing data models, which refer to objects ("resources") and their relationships. An RDF-based model can be represented in XML syntax.

  RDF Schema extends RDF and is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.

  OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

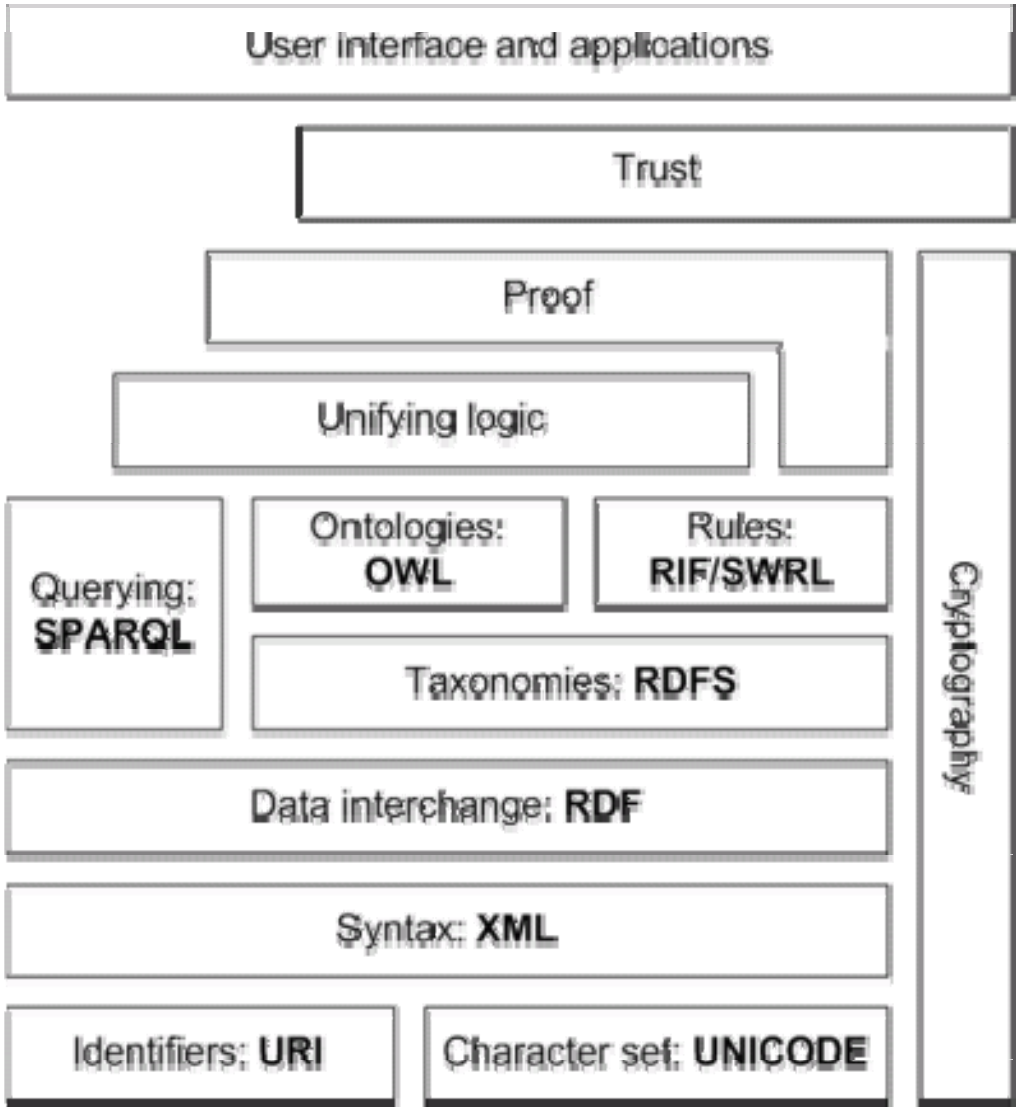
  SPARQL is a protocol and query language for semantic web data sources.

Components cont.



RDF example (documents and their authors)

Components cont.



Semantic Web Stack

6. Challenges

- Some of the challenges for the Semantic Web include vastness, vagueness, uncertainty, inconsistency and deceit. Automated reasoning systems will have to deal with all of these issues:

- **Vastness**: The World Wide Web contains at least 24 billion pages as of 2010. The SNOMED CT medical terminology ontology contains 370,000 class names, and existing technology has not yet been able to eliminate all semantically duplicated terms.

- **Vagueness**: These are imprecise concepts like "young" or "tall". This arises from the vagueness of user queries, of concepts represented by content providers, of matching query terms to provider terms and of trying to combine different knowledge bases with overlapping but subtly different concepts → Fuzzy logic

- **Uncertainty**: These are precise concepts with uncertain values. For example, a patient might present a set of symptoms which correspond to a number of different diagnoses each with a different probability → Probabilistic reasoning techniques

- **Inconsistency**: These are logical contradictions which will inevitably arise during the development of large ontologies. Deductive reasoning fails catastrophically when faced with inconsistency, because "anything follows from a contradiction“ → Defeasible reasoning and paraconsistent reasoning

- **Deceit**: This is when the producer of the information is intentionally misleading the consumer of the information → Cryptography techniques

7. Projects

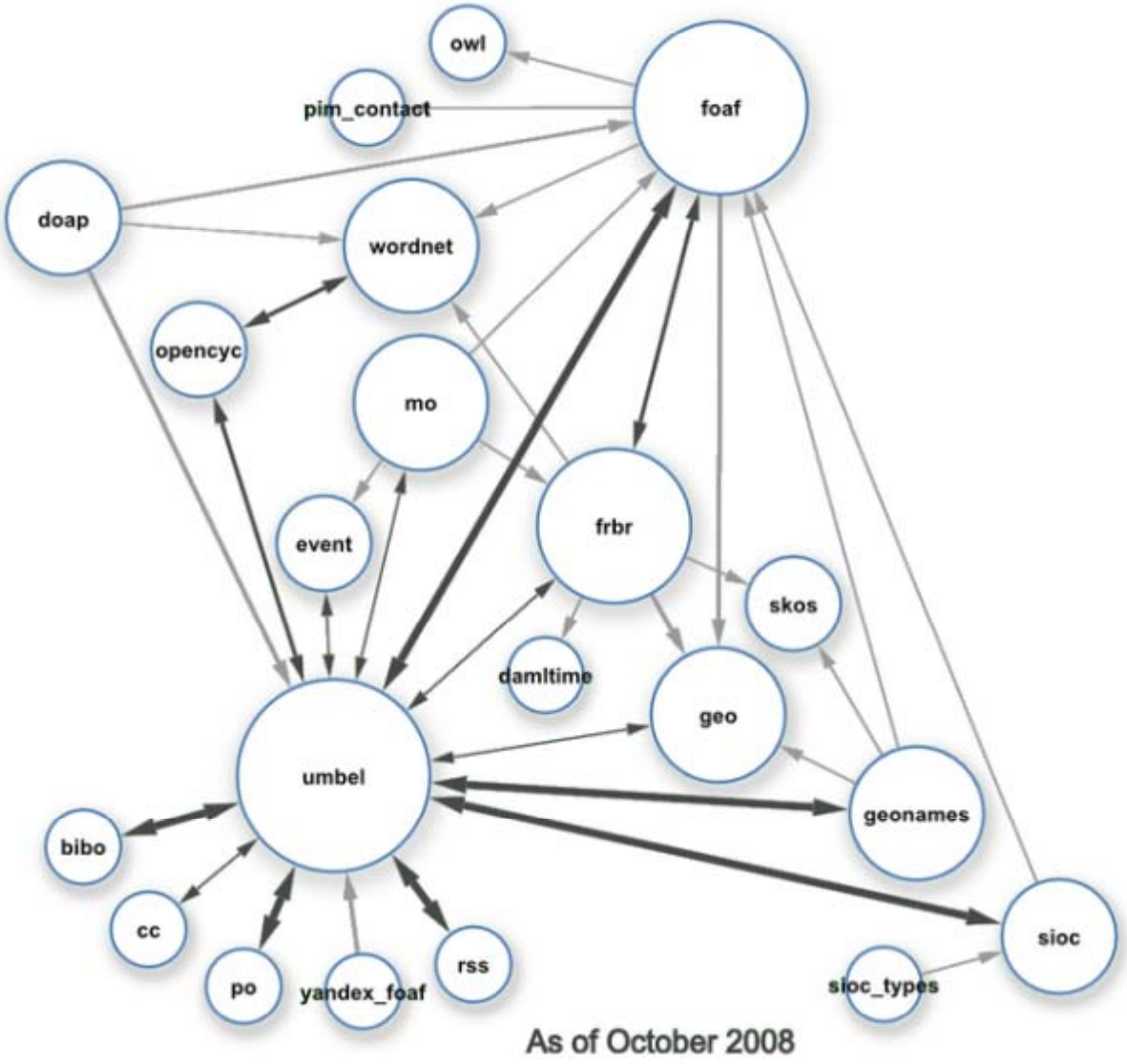
- **Dbpedia**: is an effort to publish structured data extracted from Wikipedia: the data is published in RDF and made available on the Web for use under the GNU Free Documentation License, thus allowing Semantic Web agents to provide inferencing and advanced querying over the Wikipedia-derived dataset and facilitating interlinking and extension in other data-sources.

- **FoaF**: A popular application of the semantic web is Friend of a Friend which uses RDF to describe the relationships people have to other people and the "things" around them. FOAF permits intelligent agents to make sense of the thousands of connections people have with each other, their jobs and the items important to their lives; connections that may or may not be enumerated in searches using traditional web search engines. Because the connections are so vast in number, human interpretation of the information may not be the best way of analyzing them. FOAF is an example of how the Semantic Web attempts to use the relationships within a social context.

- **GoodRelations**: is a popular vocabulary for expressing product information, prices, payment options, etc. GoodRelations has been adopted by BestBuy, Yahoo, OpenLink Software, O'Reilly Media, the Book Mashup, and many others.

Projects cont.

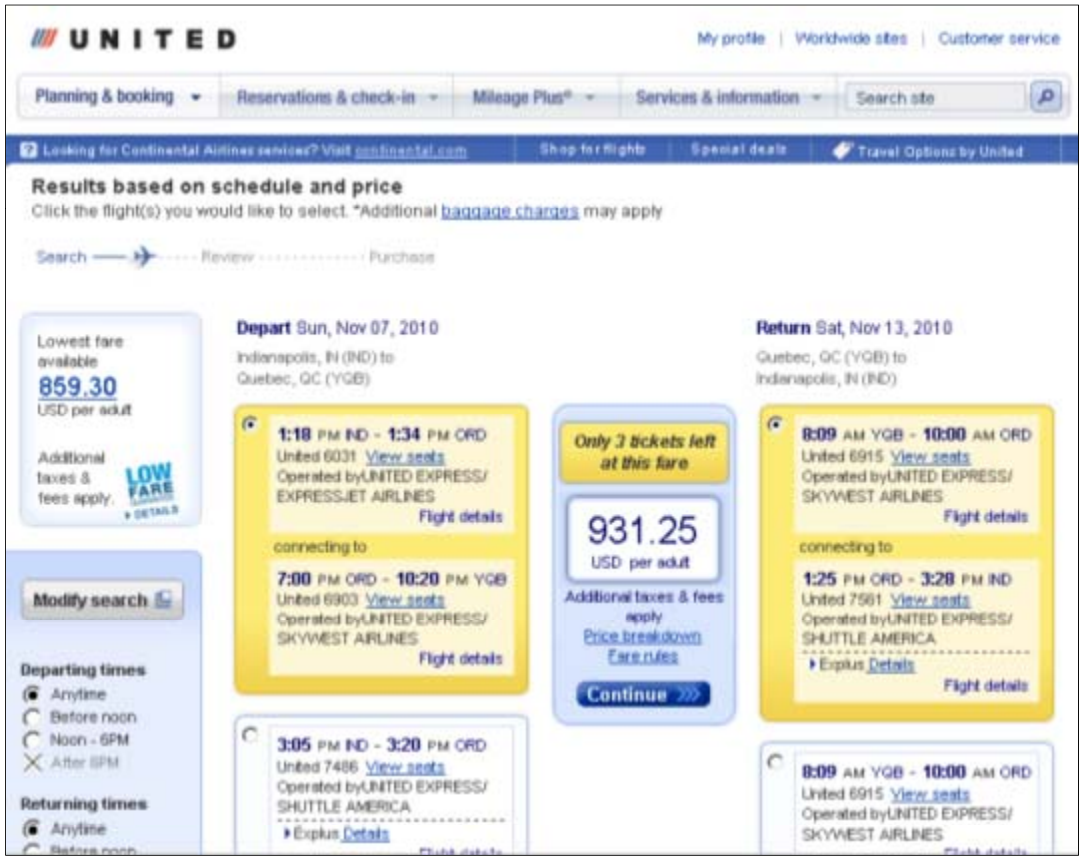
- **Linking Open Data**: is a W3C-led effort to create openly accessible, and interlinked, RDF Data on the Web. The data in question takes the form of RDF Data Sets drawn from a broad collection of data sources.



8. Example

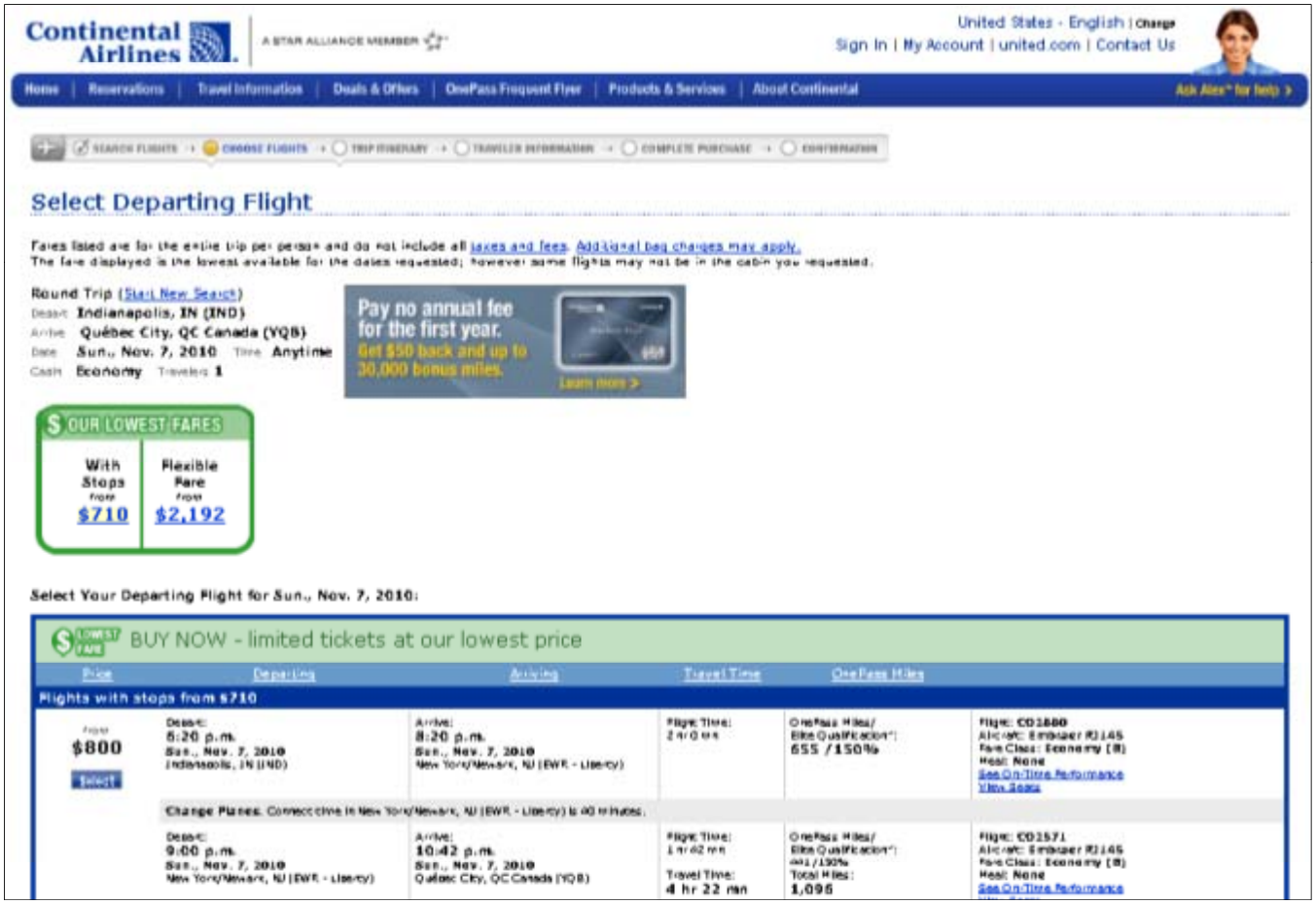
- **Example (1)**: Organizing a visit to a Quebec Hospital: Flight / Hotel / Pictures / Center

- Flight → High Cost Airline



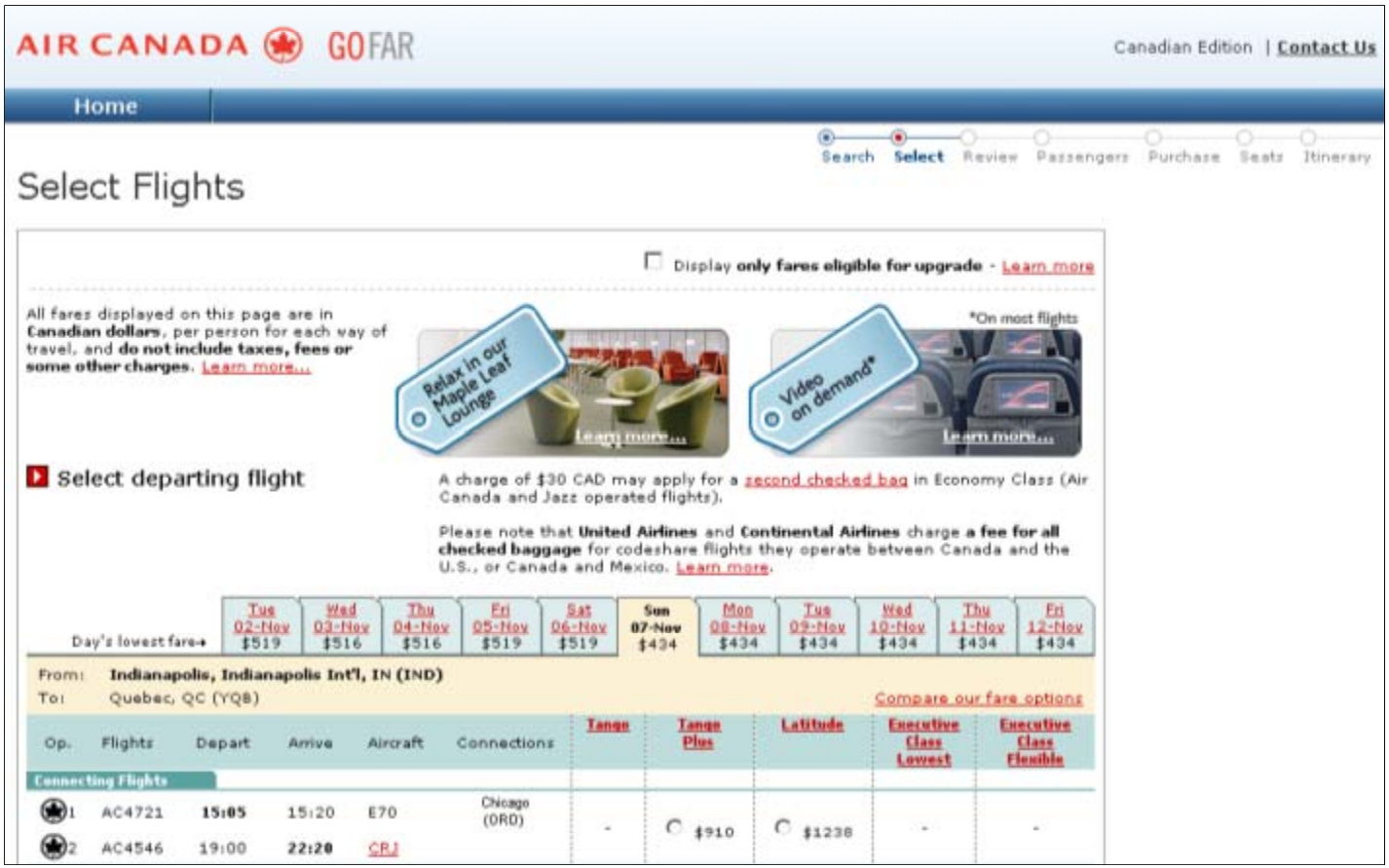
Example cont.

- Flight → Moderate Cost Airline



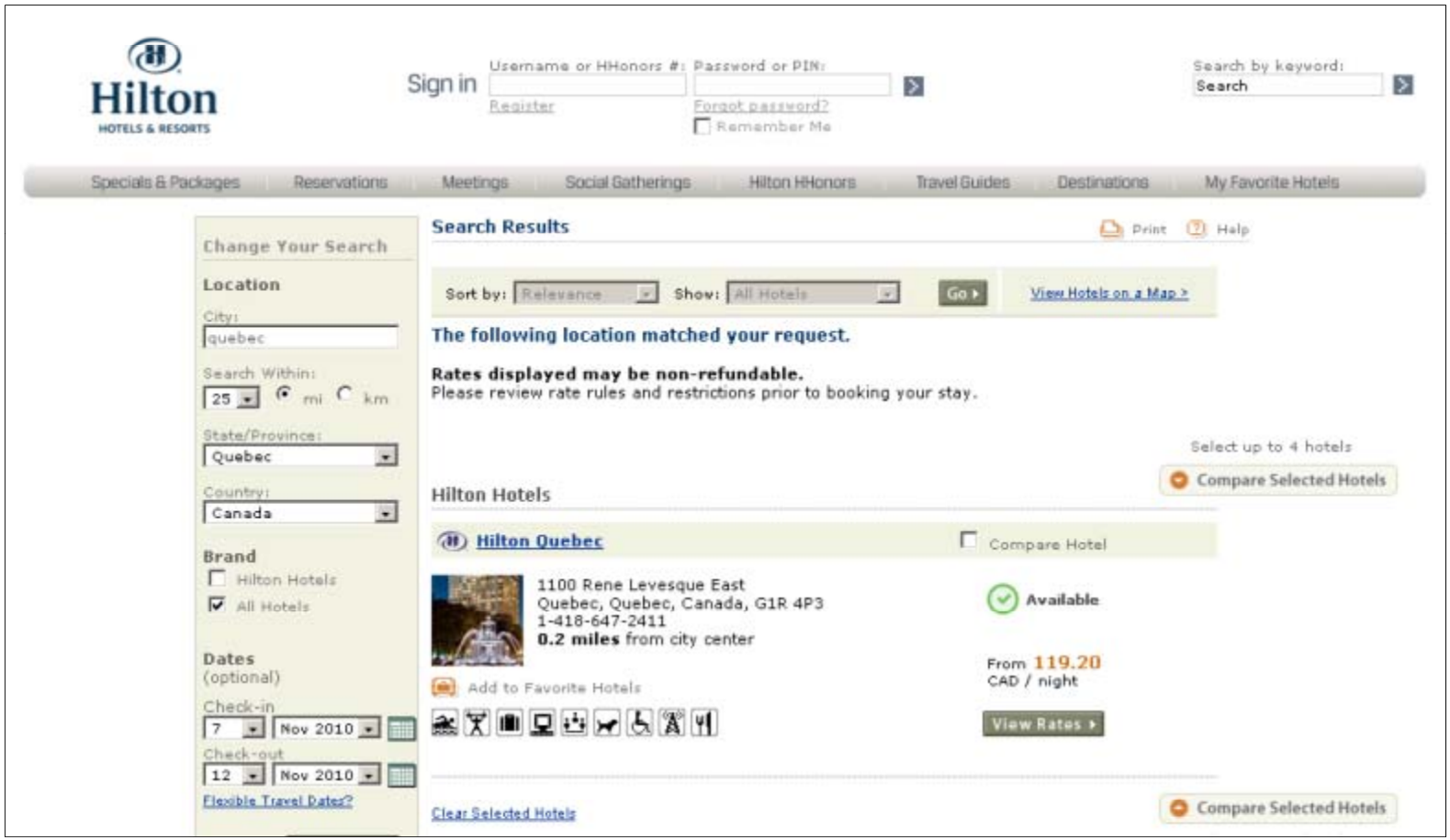
Example cont.

- Flight → Country Specific (Local) Airline



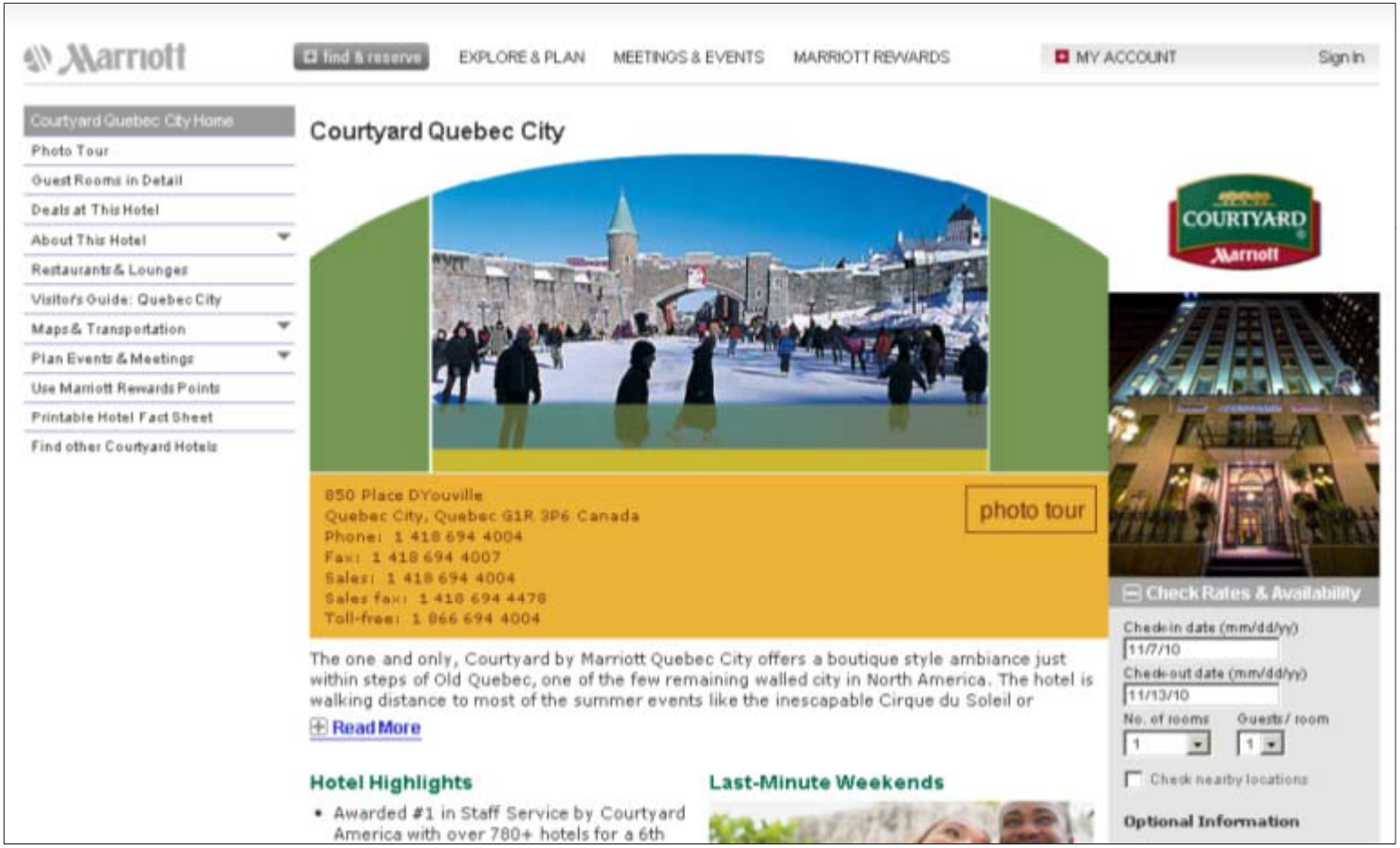
Example cont.

- Hotel → High Cost Hotel Chain



Example cont.

- Hotel → Moderate Cost Hotel Chain



Example cont.

- Hotel → Low Cost Hostel

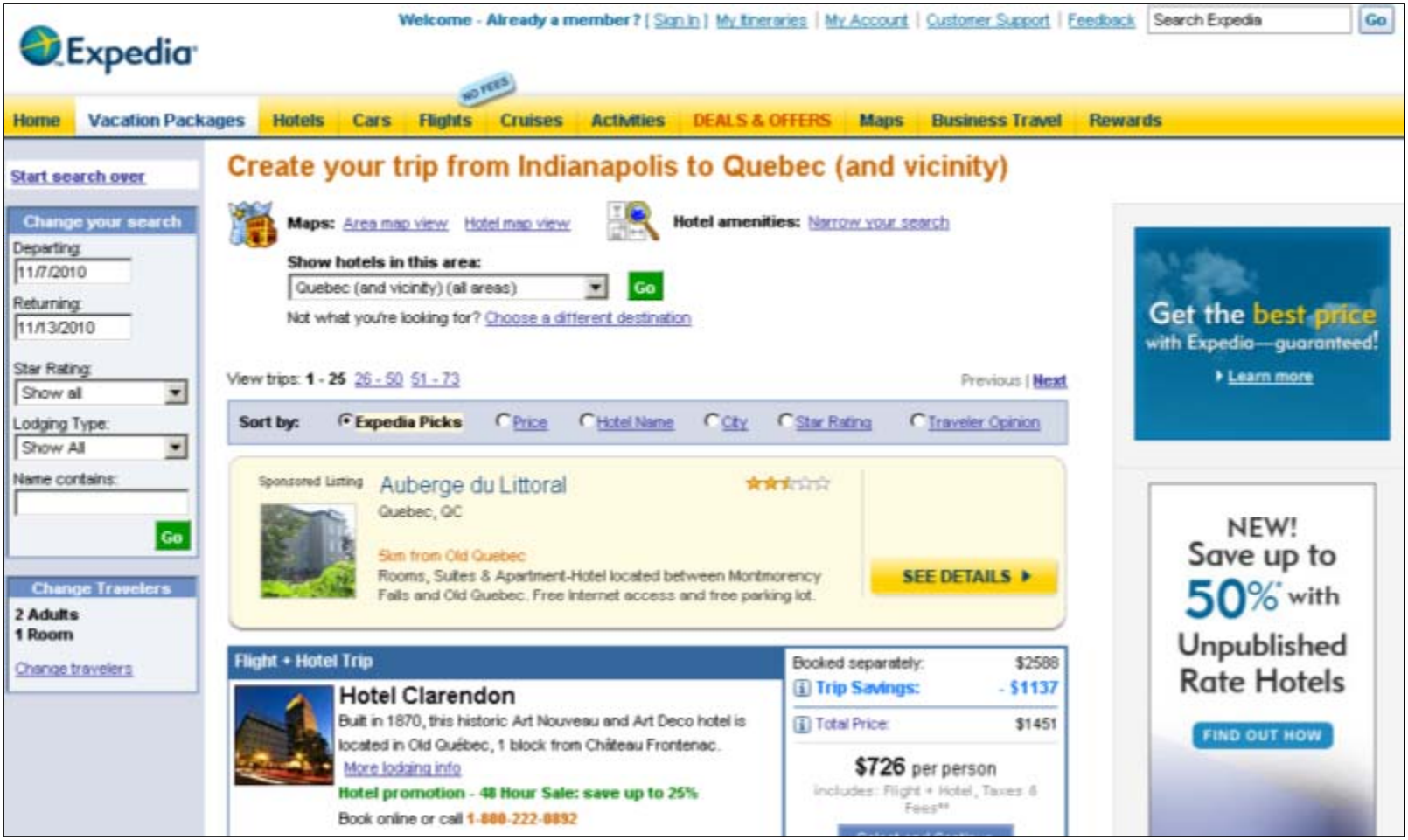
Example cont.

- Hotel → Country Specific (Local) Hotel

All in Canadian-French!

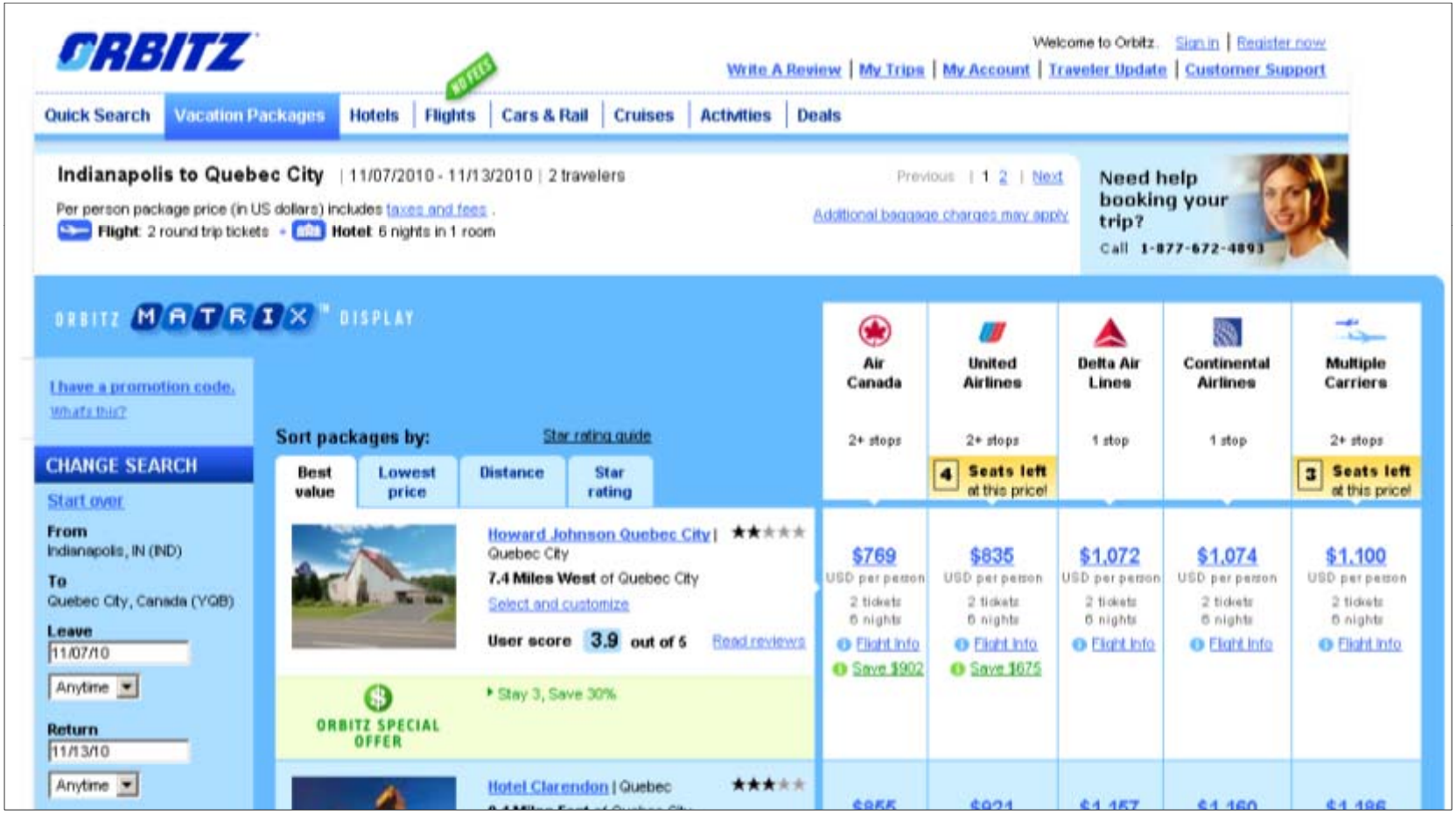
Example cont.

- Flight + Hotel → Specialized Travel Website



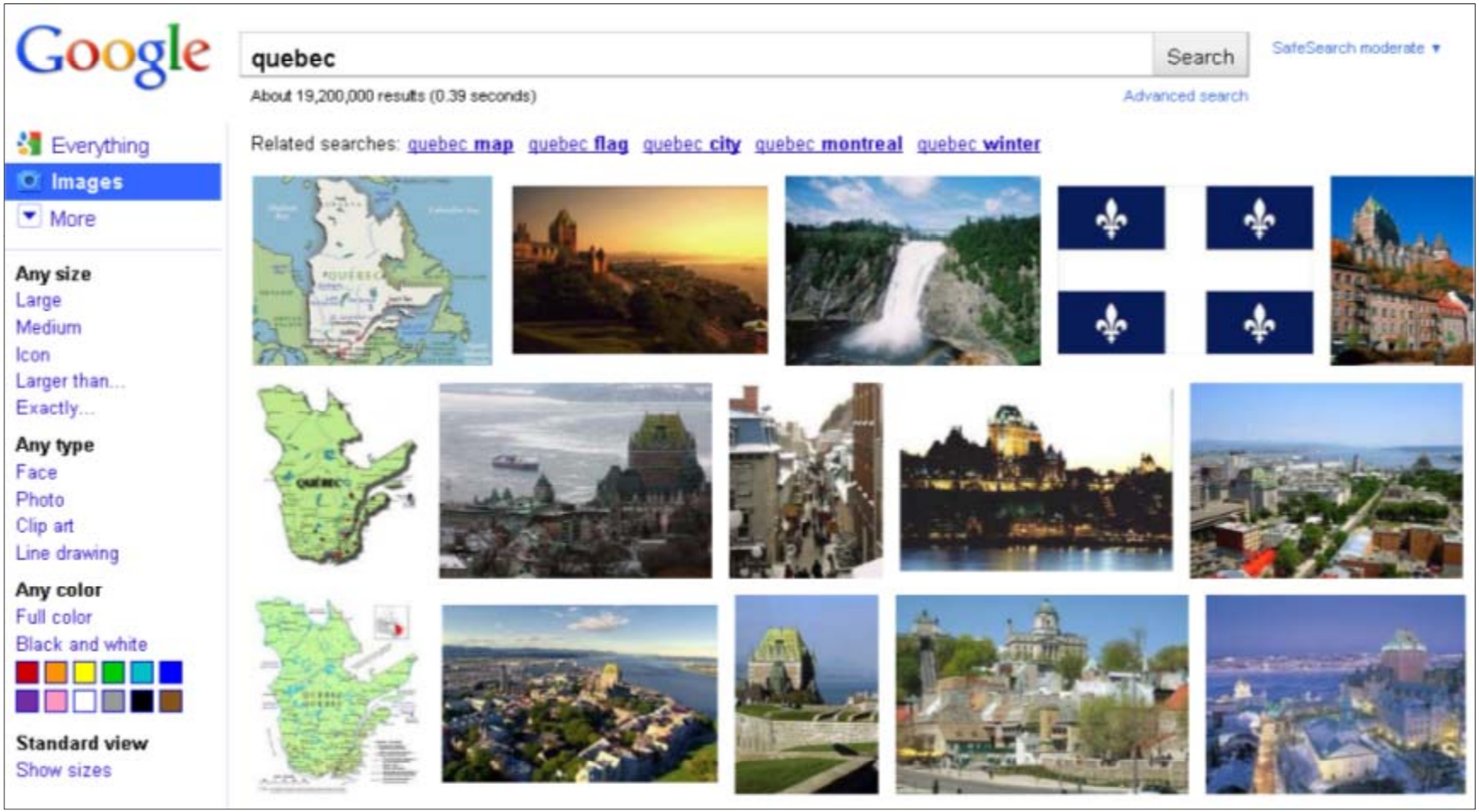
Example cont.

- Flight + Hotel → Specialized Travel Website



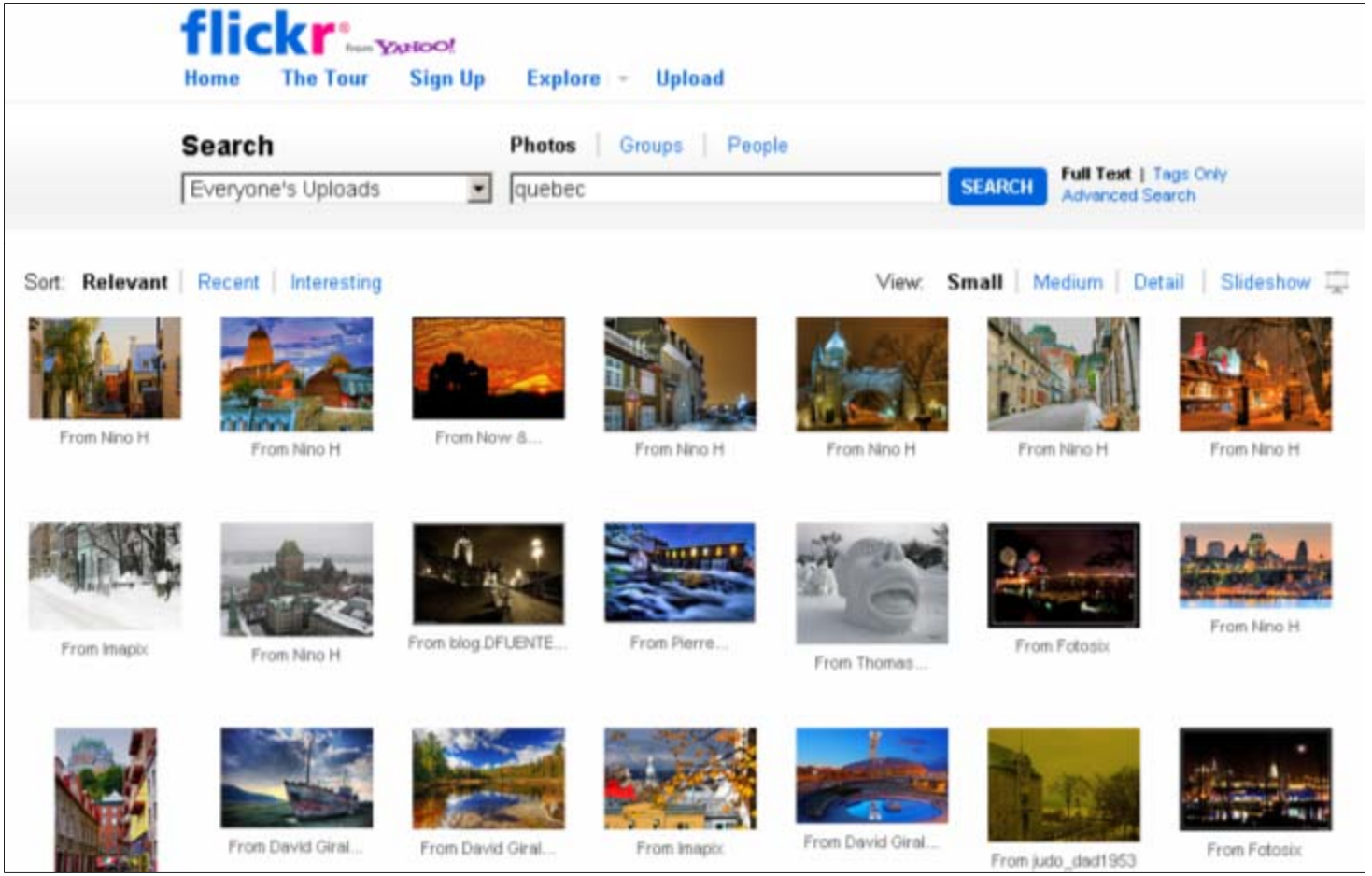
Example cont.

- Photo → Photo Search Engine (Storage) Website



Example cont.

- Photo → Photo Sharing Website



Example cont.

- Because these website are mainly human readable and not computer readable → you had to search and try a large number of websites with different languages, styles and etc → then you had to combine and integrate all those websites to find your desired information → this is a very long and exhausting process

- All of these websites have a large amount of information hidden in their backend → DBs (e.g., MySQL) and XML data islets → however you can only see the frontend (e.g., united.com) or a predefined combination of them (e.g., expedia.com)

- Goal → We like to be able to link to data independently of their presentation and use it the way we want not the way that the designer wanted to present → agents (e.g., web programs) should be able to interpret data from various website the way we want → **Web of Data**

- Mashup Websites → collect data from other websites (that are not necessarily web semantic compatible but provide their information via web services) and offer the information in one spot. Indeed, Mashup websites give us a sneak peak of what the true power of Web of Data can be.

Example cont.

- **Example (2)**: Bookstore Database En, Fr, and Wikipedia Integration → (a) Map the various data onto an abstract data representation (b) Merge the resulting representations (c) Start making queries on the whole.

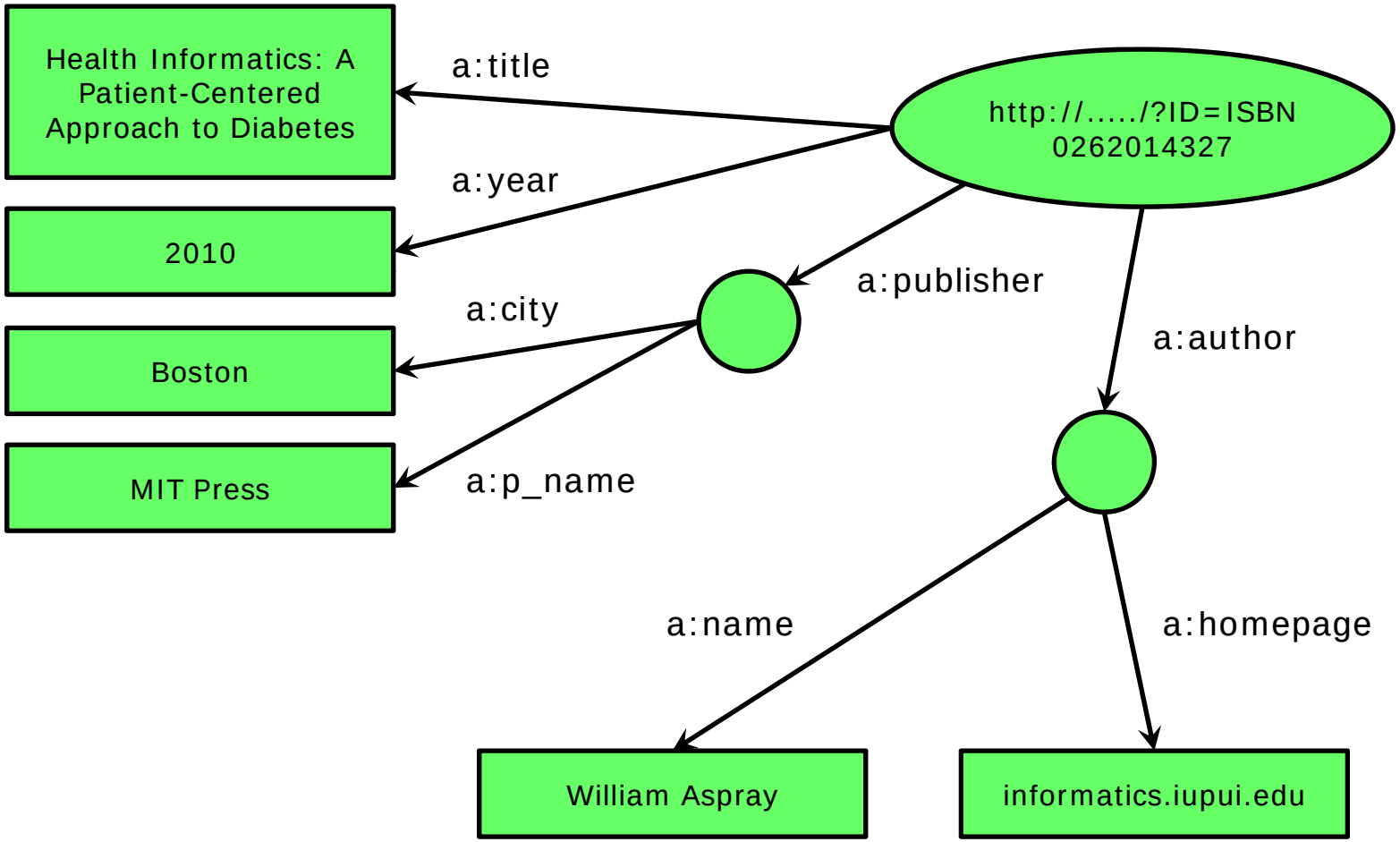
- Sample En Bookstore Dataset (e.g., MySQL)

ID	Author	Title	Publisher	Year
ISBN0262014327	id_xyz	Health Informatics: A Patient-Centered Approach to Diabetes	id_qpr	2010

ID	Name	Home Page
id_xyz	Aspray, William	http://informatics.iupui.edu

ID	Publ. Name	City
id_qpr	The MIT Press	Boston

Example cont.



Example cont.

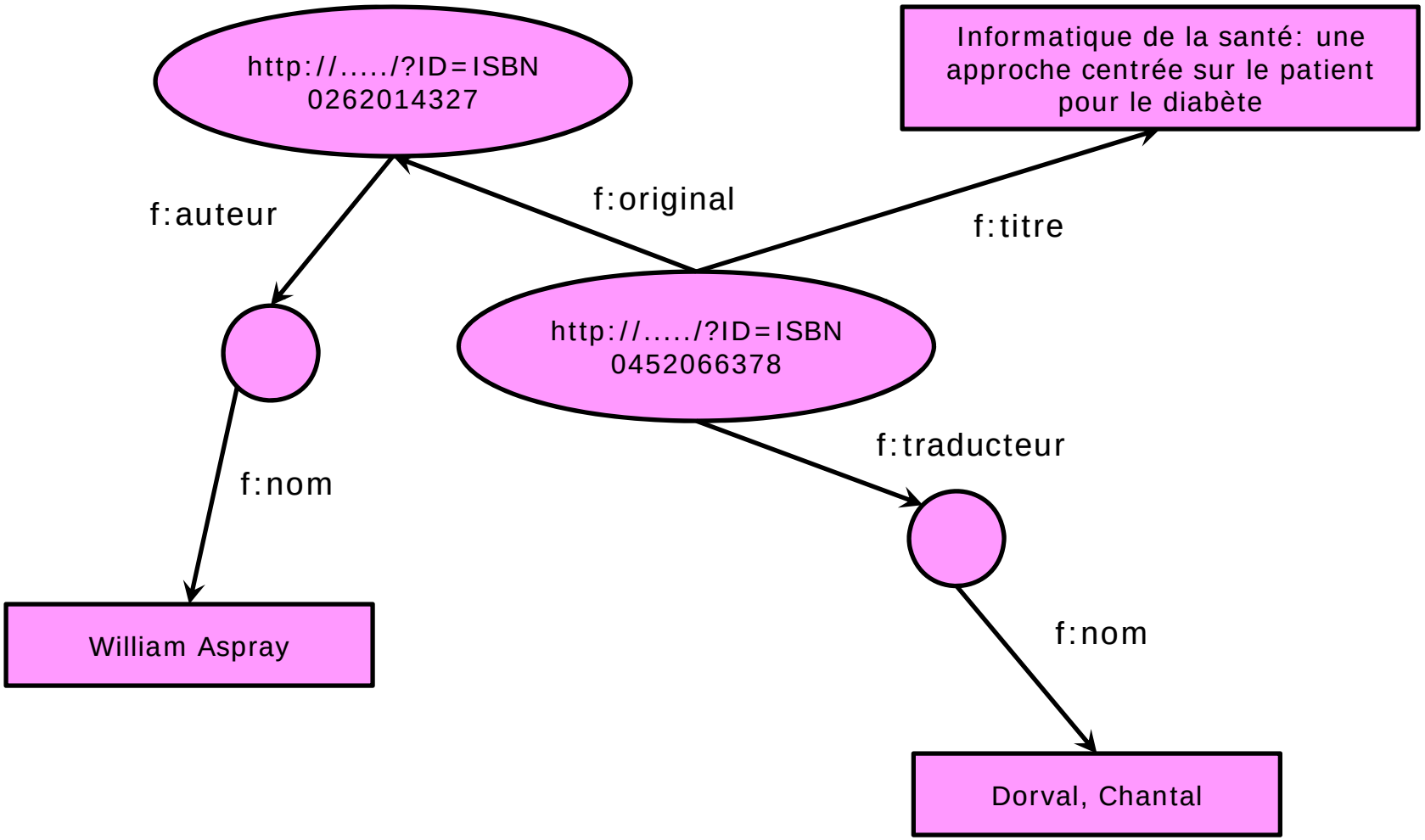
- Sample Fr Bookstore Dataset (e.g., Excel)

	A	B	D	E
1	**ID**	**Titre**	**Traducteur**	**Original**
2	ISBN0452066378	Informatique de la santé: une approche centrée sur le patient pour le diabète	A13	ISBN0262014327
3				
4				
5				

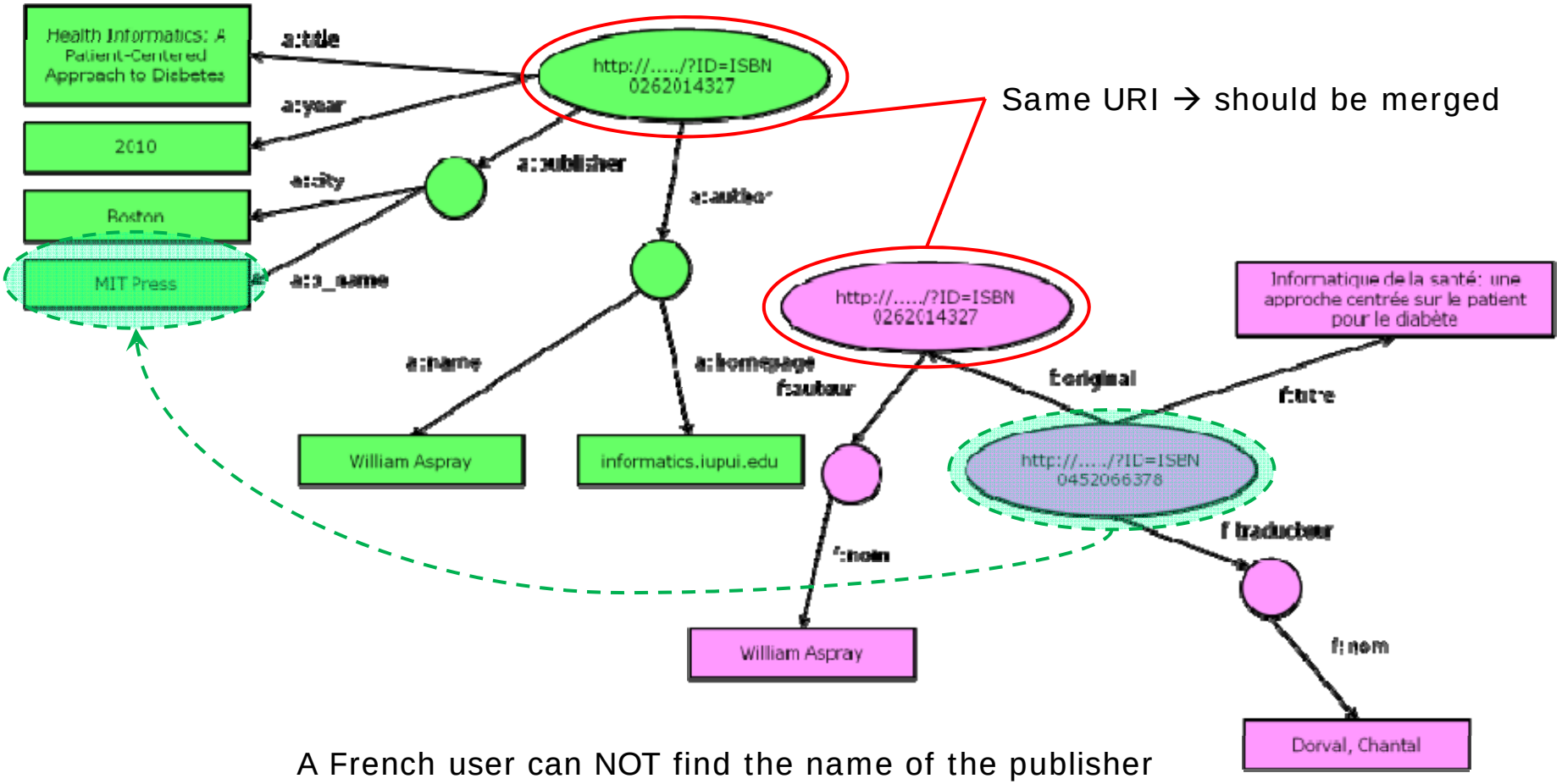
6	**ID**	**Auteur**
7	ISBN0262014327	A12
8		
9		
10		

11	**Nom**
12	Aspray, William
13	Dorval, Chantal

Example cont.

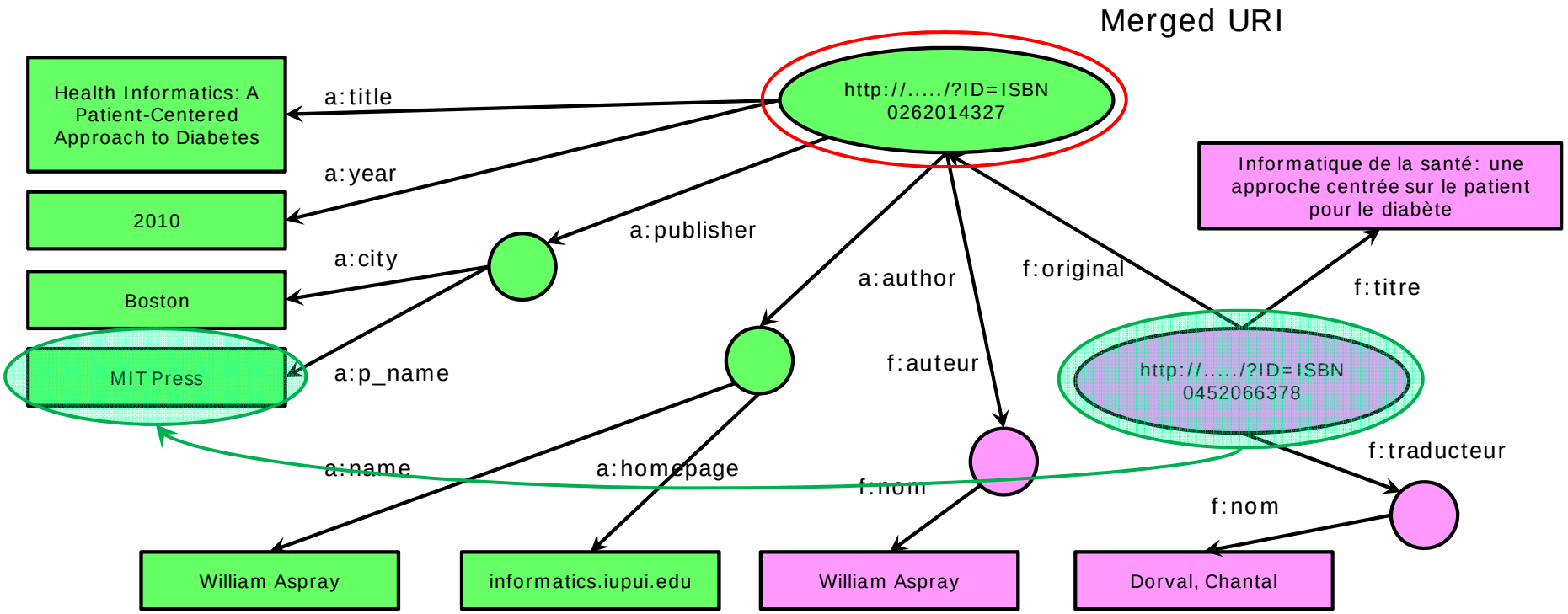


Example cont.



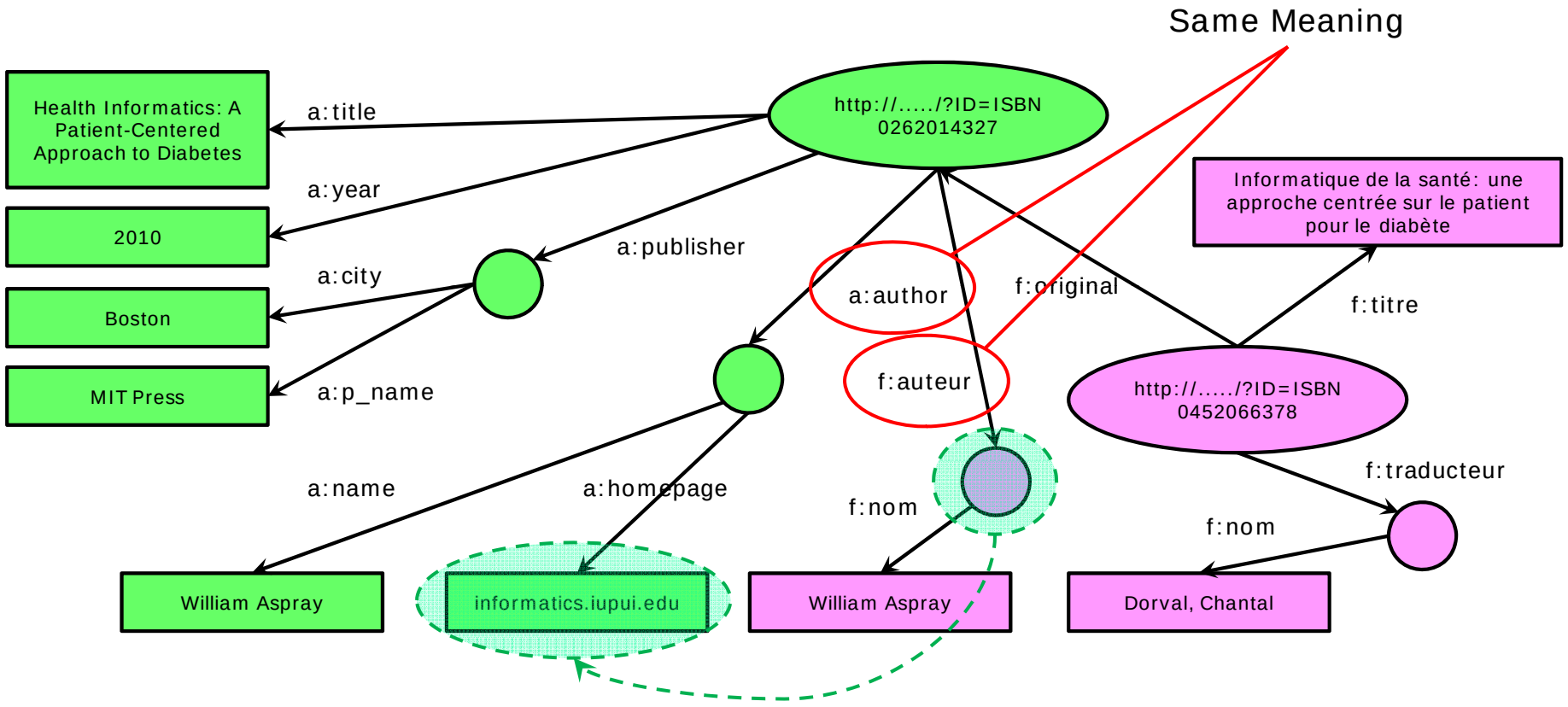
A French user can NOT find the name of the publisher

Example cont.



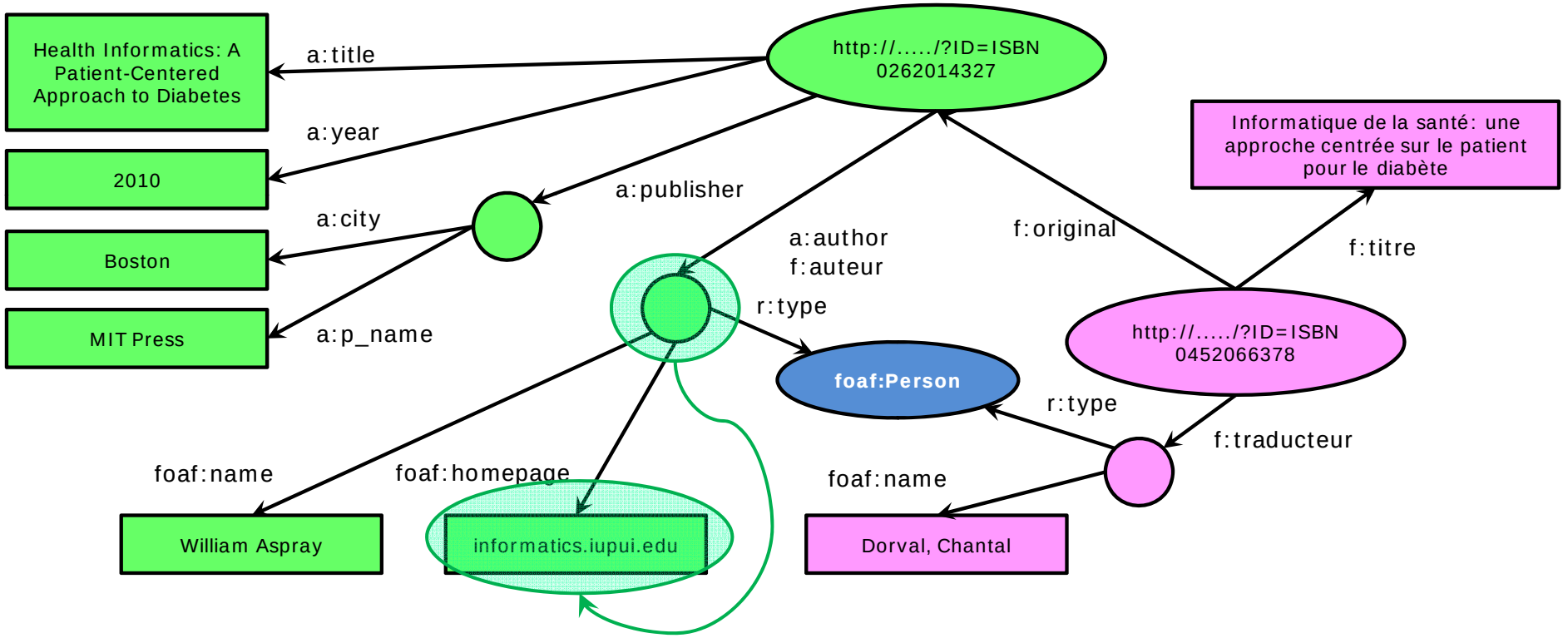
A French user can find the name of the publisher

Example cont.



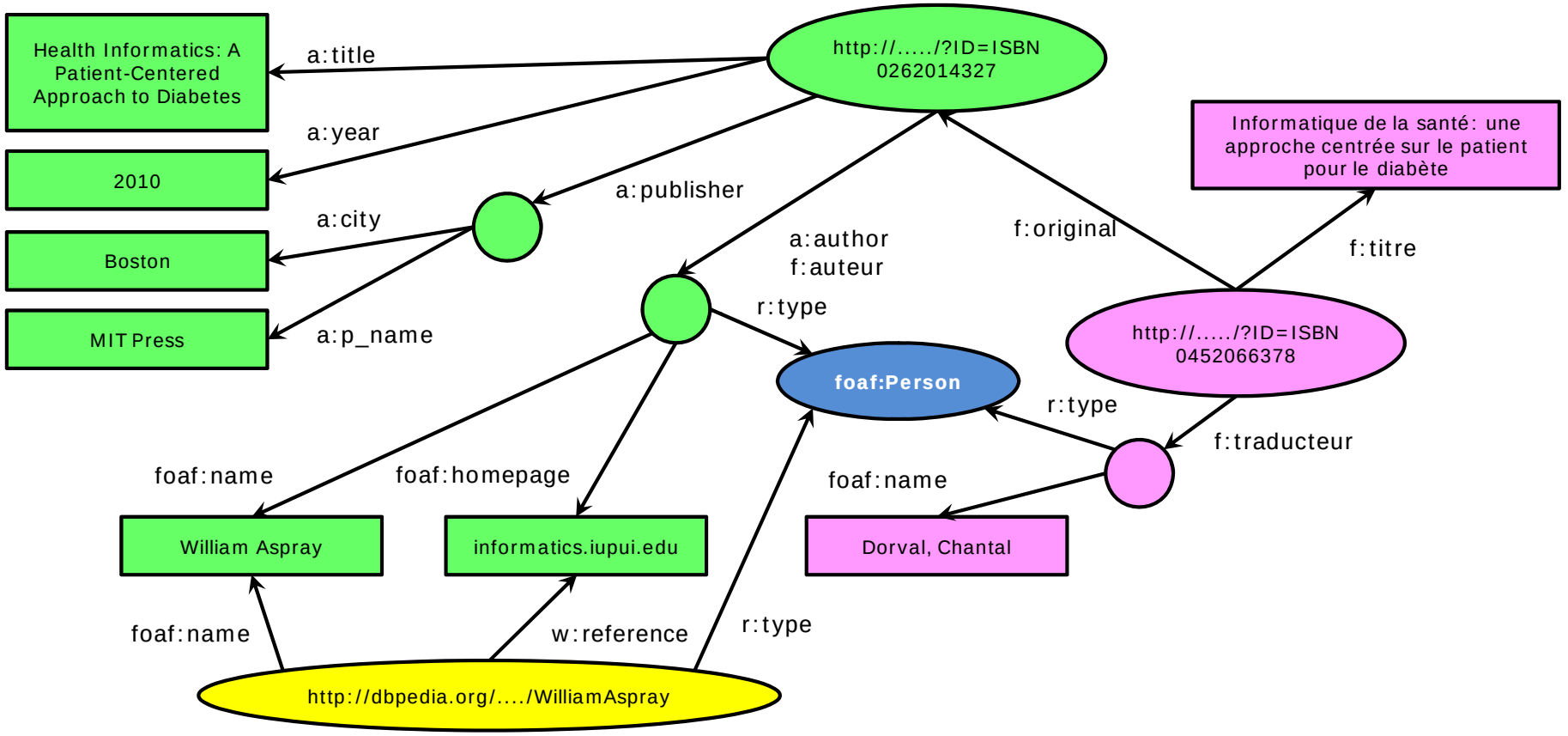
A French user can NOT find the author's website

Example cont.



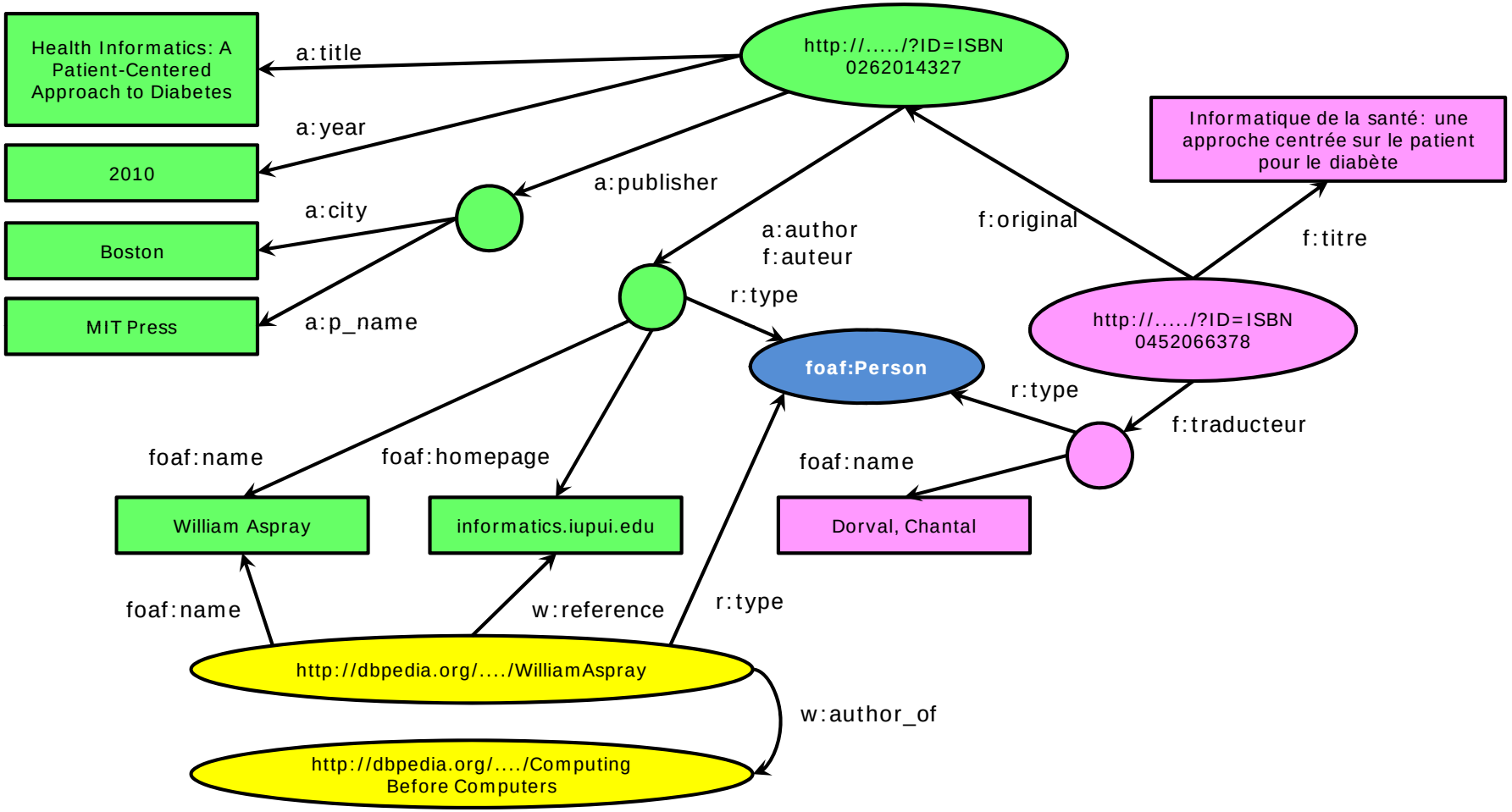
A French user can find the author's website

Example cont.



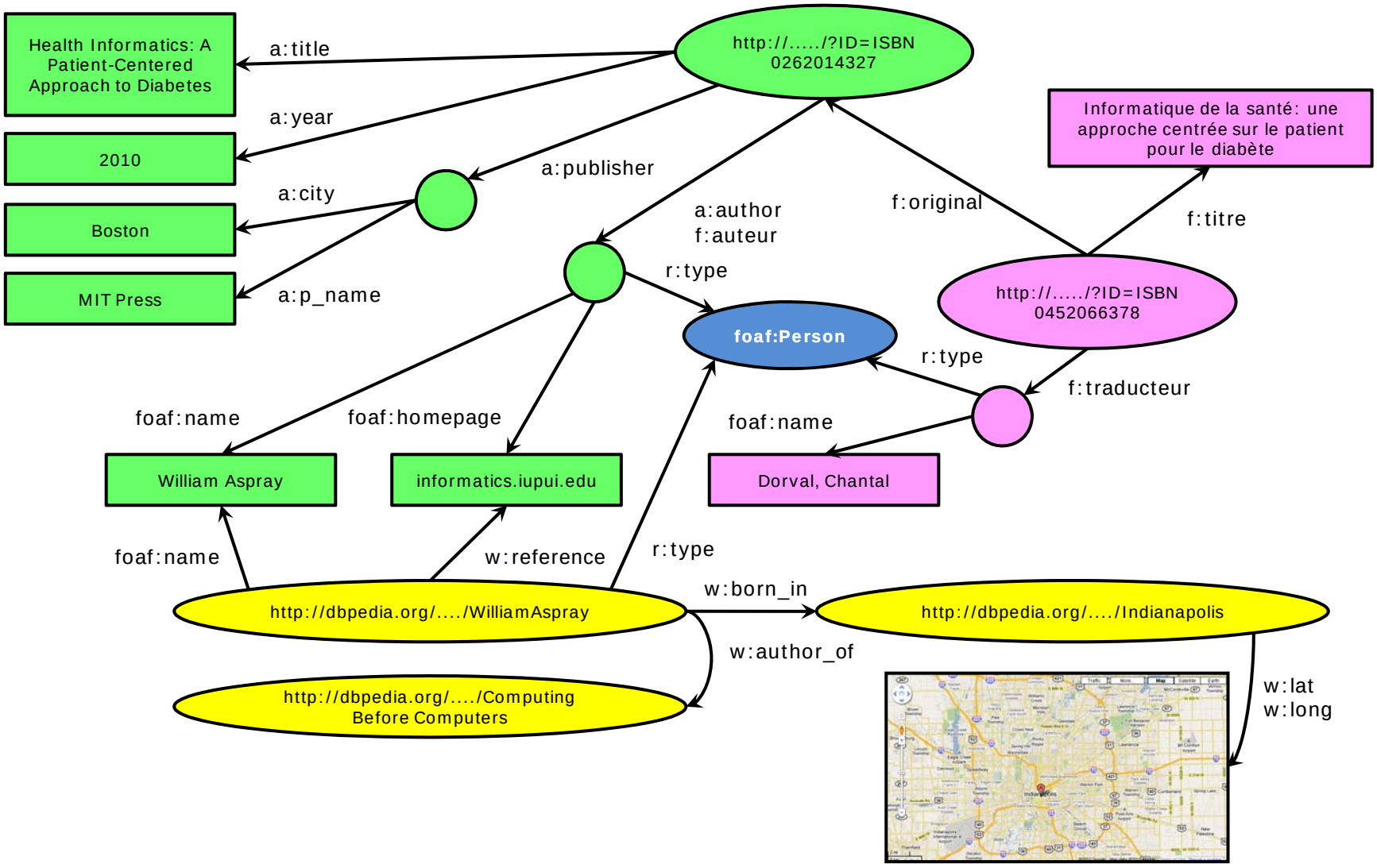
Integration into Wikipedia via DBpedia

Example cont.



Further integration into Wikipedia via DBpedia

Example cont.

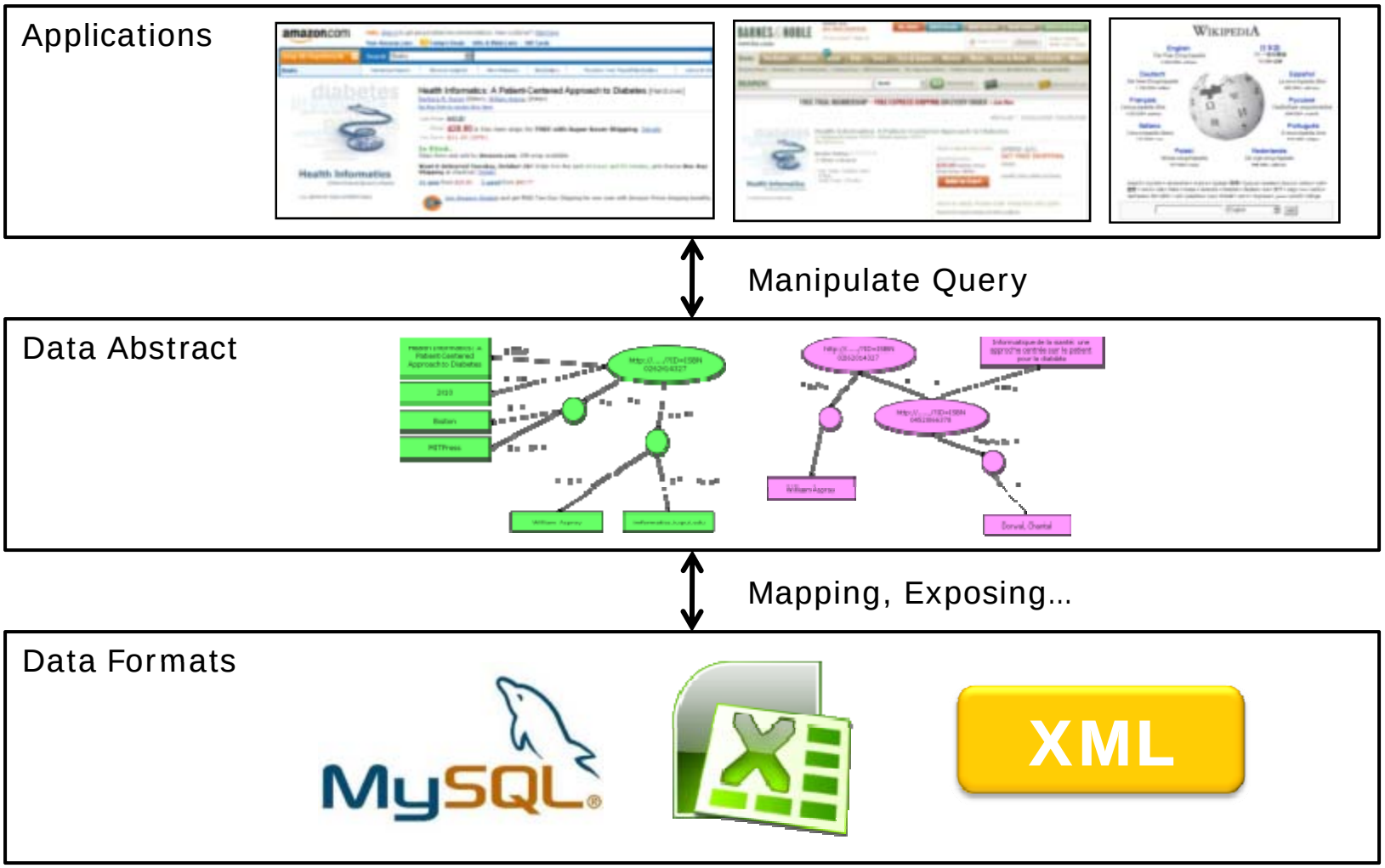


Further integration into Wikipedia via Dbpedia + Google Map

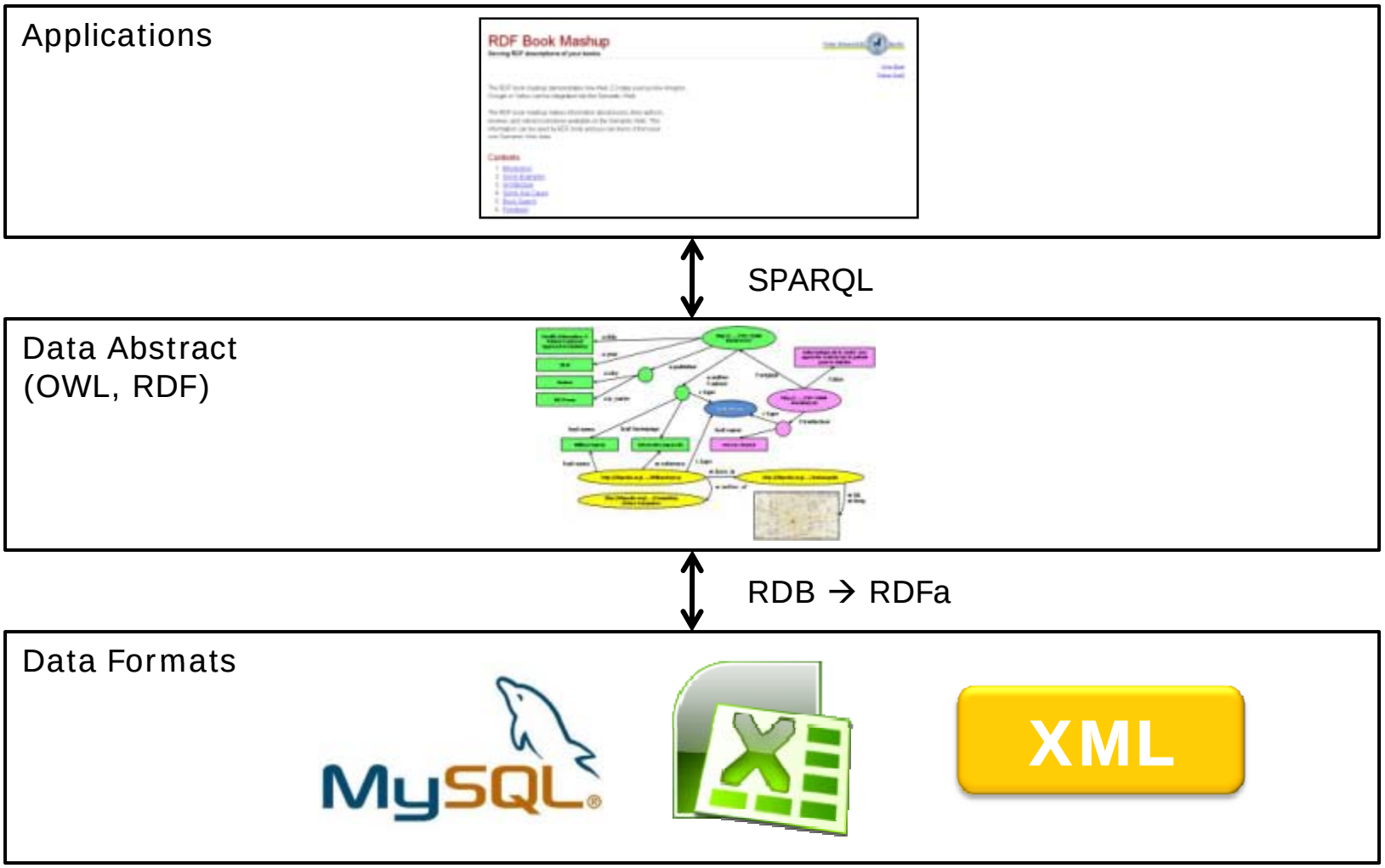
Example cont.

- What did we do?
  - We combined different datasets that
    - are somewhere on the web
    - are of different formats (mysql, excel sheet, XHTML, etc)
    - have different names for relations
  - We could combine the data because some URI-s were identical (the ISBN-s in this case)
  - We could add some simple additional information, using common terminologies that a community has produced
  - As a result, new relations could be found and retrieved

Example cont.



Example cont.



Summary

1. Definition
2. Purpose
3. Relationship to The Hypertext Web
4. Skeptical Reactions
5. Components
6. Challenges
7. Projects
8. Example