Applying Semantic Web Technologies for Improving the Visibility of Tourism Data

Fayrouz Soualah-Alila, Cyril Faucher, Frédéric Bertrand,
Mickaël Coustaty, Antoine Doucet

L3I laboratory
La Rochelle University
Michel Crépeau avenue, 17000 La Rochelle, France
{fayrouz.soualah-alila, cyril.faucher, fbertran, mickael.coustaty, antoine.doucet}@univ-lr.fr

ABSTRACT

Tourism industry is an extremely information-intensive, complex and dynamic activity. It can benefit from semantic Web technologies, due to the significant heterogeneity of information sources and the high volume of on-line data. The management of semantically diverse annotated tourism data is facilitated by ontologies that provide methods and standards, which allow flexibility and more intelligent access to on-line data. This paper provides a description of some of the early results of the Tourinflux project which aims to apply semantic Web technologies to support tourist actors in effectively finding and publishing information on the Web.

Categories and Subject Descriptors

E.2 [**Data**]: Data Storage Representations; D.2.12 [**Software**]: Software Engineering—*Interoperability*

General Terms

Tourism, Semantic Web

Keywords

Tourinflux, TourInFrance, Ontology, Schema.org, Mapping

1. INTRODUCTION

1.1 Context

The tourism is viewed as information intensive industry and a highly dynamic and changing domain where information plays an important role for decision and action making [5]. Tourism represents a significant part of the global economy, more than 9% of the global GDP (US $7 trillion) [10]. It becomes increasingly difficult for policy makers, on the one hand to guide choices in order to boost and make their territory attractive, and on the other hand to analyze the comments and the vision of the territory by users. As the

World Wide Web has changed people's daily life, it has significantly influenced the way of information gathering and exchanging in the tourism area, with the intensive use of social networks and Web sites specialized in e-tourism (TripAdvisor, Booking.com, etc.). Along with the Web technologies, tourism actors such as tour operators need knowledge about travel, hotels, markets, etc. Indeed, information like hotel prices, hotel localization, offers, market tendencies are used to improve their service quality by enhancing employee's knowledge about customer's preferences. Other actors such as experts from tourism industry are also making use of the Web to promote their territory, and so to attract more visitors.

However, the gaps of the existing Web technologies arise two main topics:

First, tourism domain is characterized by a significant heterogeneity of information sources and by a high volume of on-line data. Data related to tourism is produced by different experts (travel agents, tourist offices, etc.) and by visitors, constituting semantically heterogeneous data, often incomplete and inconsistent. This data could be composed of information related to tourism objects (hotels, concert, restaurant, etc.), temporal information or opinions. There already exist different taxonomies and catalogues which are designed and used internally by tourism actors to help them to manage heterogeneous tourism data. Efforts are now made to generate standards to facilitate inter and intra tourism data exchange.

Second, the information contained on Web pages are originally designed to be human-readable, and so, most of information currently available on the Web are kept in large collections of textual documents. As the Web grows in size and complexity, there is an increasing need for automating some of the time consuming tasks such as information searching, extraction and interpretation. Semantic Web technologies are an answer to this issue by proposing some novel and sophisticated question answering systems. In this way, ontologies provides a formal framework to organize data and to browse, search or access this information [2].

1.2 Tourinflux

This work takes place under the Tourinflux[1] project which aims at providing the tourism industry with a set of tools (1) allowing them to handle both their internal data, and the information available on the Web, and (2) allowing to improve the display information available about their terri-

---

[1] http://tourinflux.univ-lr.fr/

tory on the Web. This should allow them to improve their decision process and subsequently, their effectiveness.

This paper presents some preliminary results of Tourinflux project. The paper aim to describe the development of an ontology based model for tourism domain. It is aimed primarily to present some aspects of knowledge management: Section 2 reviews the related ontologies developed in the previous projects. In particular, we describe how an ontology can be used to improve the quality of users Web search and data publication. Section 2 also provides description of current tourism standards and a brief background on a particular ontology: the Schema.org[2] model. In an early stage of our project, a partial ontology called TIFSem (Semantic TourInFrance) for the tourism was created, using Protégé [3] and the Web Ontology Language [4] (OWL). A partial view of TIFSem ontology developed will be presented in section 3 with some example of queries. We notice that this work is in progress; we are still gathering new concepts for its taxonomy and new axioms in order to validate them with experts of the domain. Finally, conclusions and outlook are provided in section 4.

2. STATE OF ART

2.1 Tourism ontologies

In the domain of tourism, several researches have focused on the design of ontologies. Several available tourism ontologies show the current status of the efforts:

The OTA (OpenTravelAlliance) specification defines XML Schema documents corresponding to events and activities in various travel sectors [7].

The Harmonise ontology is specialised to address interoperability problems focusing on data exchange [4]. Harmonise is based on mapping different tourism ontologies by using a mediating ontology. This central Harmonise ontology is represented in RDF and focuses mainly on tourism events and accommodation types.

The Hi-Touch ontology models tourism destinations and their associated documentations [6]. The ontology classifies tourist objects, which are linked together in a network by semantic relationships. The top-level classes of the Hi-Touch ontology are documents (any kind of documentation about a tourism product), objects (the tourism objects themselves) and publication (a document created from the results of a query).

The QALL-ME ontology provides a model to describe tourism destinations, sites, events and transportation [8] and aims at establishing a shared infrastructure for multilingual and multimodal question answering. The Tourpedia catalogue describes places (points, restaurants, accommodations and interests) related to different locations (Amsterdam, London, Paris, Rome, etc.) [3].

These models focus on different areas of tourism domain but there does not exist one single ontology which matches all the needs of different tourism related applications. We then propose in section 3 an ontology to globally describe tourism information mixing heterogeneous content. Concepts included in the defined ontology will allow us to annotate information sources on tourism. Then the enriched information can be used: (1) from the user side, to match

---

tailored package holidays to client preferences and (2) from tourism experts' side, to analyze and better manage on-line data about their territory.

2.2 Tourism standards

The interoperability of Tourism Information Systems (TIS) is a major challenge for the development of tourism domain. Several national, European and international institutional initiatives have proposed different standards to meet the specific needs of tourism professionals. In France, professionals have adopted the TourInFrance standard (TIF). Since its creation in 1999 by the tourism ministry, TourInFrance is used today by more than 3000 tourist offices in France, by Departmental Tourist Committees (DTC) and by different tour operators, to facilitate data exchange between these different actors. In 2004, the TourInFrance Technical Group (TIFTG) approved the new version of the standard, TIF V3. In this version, the standard has evolved towards XML technologies to facilitate publishing information on the Web and information exchange between systems, and is accompanied by several thesaurus. The last version of TIF can be downloaded from the following Web site: http://tourinfrance.free.fr/. Since 2005, this standard stopped evolving. As a result, tourism professionals have adapted the standard to their own needs and proposed their own evolution in an unorganized way, producing TIS that are not really interoperable among themselves and that cannot directly be shared using international standard. Accordingly, the exploitation of tourism information is trapped in its own territory, and so it is impossible to aggregate these information.

To illustrate this evolution of TIF we show in the following example how the standard was transformed. The figures 1 and 2 show different syntaxes of TIF to describe an address; data represented in the figure 1 respect the standard TIF V3, while data represented in the figure 2 is structured under a derived version of TIF (different tags syntaxes, new tags added, etc.).

To overcome these limitations we want to evolve the TIF standard to share the knowledge it represents and to ensure data interoperability, by applying the concept of ontology to represent the standard terminology [1].

```xml
▼<tif:Adresses>
  ▼<tif:DetailAdresse>
    <tif:CodePostal>10000</tif:CodePostal>
    <tif:Commune code="10387">Troyes</tif:Commune>
    <tif:Adr1>18 rue Linard Gonthier</tif:Adr1>
    ▼<tif:Personnes>
      ▼<tif:DetailPersonne libelle="Moyens de communication" ty
        ▼<tif:MoyensCommunications>
          ▼<tif:DetailMoyenCom libelle="Téléphone" type="04.02.
            <tif:Coord>0325431111</tif:Coord>
          </tif:DetailMoyenCom>
          ▼<tif:DetailMoyenCom libelle="Télécopieur /fax" type=
            <tif:Coord>0325431043</tif:Coord>
          </tif:DetailMoyenCom>
          ▼<tif:DetailMoyenCom libelle="Site web (URL)" type="0
            <tif:Coord>http://www.maisonderhodes.com</tif:Coor
          </tif:DetailMoyenCom>
          ▼<tif:DetailMoyenCom libelle="Mél" type="04.02.04" xm
            <tif:Coord>message@maisonderhodes.com</tif:Coord>
          </tif:DetailMoyenCom>
        </tif:MoyensCommunications>
      </tif:DetailPersonne>
    </tif:Personnes>
    <tif:Adr2/>
  </tif:DetailAdresse>
</tif:Adresses>
```

Figure 1: Data respecting the standard TIF V3

```xml
▼<adresse type="produit">
    <civilite/>
    <raisonSociale/>
    <prenom/>
    <nom>"P'TIT DEJ - HOTEL Ile de Ré"</nom>
    <fonction/>
    <ligne1>12 rue de la Terre Rouge</ligne1>
    <ligne2/>
    <ligne3/>
    <cp>17740</cp>
    <commune codeInsee="17360">STE MARIE DE RE</commune>
    <paysMonde iso="FR">FRANCE</paysMonde>
    <tel1>05 46 30 33 20</tel1>
    <tel2/>
    <fax/>
    <email>hotel@ptitdej-hotel-iledere.com</email>
    <web>http://www.ptitdej-hotel-iledere.com</web>
    <facebook/>
    <twitter/>
    <googlePlus/>
    <tumblr/>
    <linkedIn/>
    <viadeo/>
    <instagram/>
</adresse>
```

**Figure 2:** Data structured under a derived version of TIF

2.3 Semantic Web standards

TIF standard is unable to easily share and interoperate with global Web standards. Since few years, an initiative led by Bing, Google and Yahoo tends to become a de-facto norm to easily share semantic content. Launched in 2011, Schema.org aims to create and to support a common set of schemas for structured data markup Web pages. The purpose of Schema.org is to provide a collection of schemas (HTML tags) that webmasters can use to markup HTML pages in ways recognized by major search providers, and that can also be used for structured data interoperability (RDFa, JSON-LD, etc.). Upon using these tags in a website, search engines can understand the intention of using resource (text, image, video) in that website better; which in turn would return better search results when a user is looking for a specific resource through a search engine [9]. Schema.org is a very large vocabulary counting hundreds of terms from multiple domains (the full specification of Schema.org is available at https://schema.org/docs/full.html).

In our approach we want to spread enriched semantic tourist data that can be easily indexed by search engines. One solution would be to use directly the Schema.org ontology which provides good indexing by search engines and which is easy to implement. The disadvantages of this solution is that we lose precision in the tourist data of the TIF model, in addition, using directly Schema.org can cause economic problems because a large number of tourist offices already use TIF or derived versions. A second solution would be to realize an ontology by matching terms of TIF with terms of Schema.org by using OWL relations, and work with Schema.org community to extend the schema, either formally by adding new terms or informally by defining how Schema.org can be combined with some additional vocabulary terms. We can thus produce more accurate information than the terms of Schema.org, in addition with a TIF/Schema.org correspondence, producing an interoperable model becomes possible from XML TIF sources.

This method is more complex to implement because it requires to connect each term of TIF with a terms of Schema.org. The difficulty is not the transformation of XML TIF to OWL, but searching correspondence between the terms of TIF and those of Schema.org, and seeking what is lacking in each model. We suggest to consider the second solution,

which presents a better compromise.
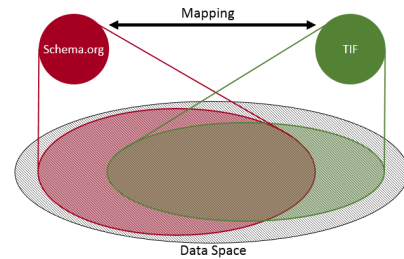
Figure 3: Schema.org/TIF mapping

We consider the Schema.org as the extension core of the TIF standard to improve the online visibility of touristic service providers. Schema.org includes many of the TIF terms. Of course not all of the Schema.org terms are relevant for the tourism domain. The relevant terms are those that belong to the categories Hotels, Food Establishments, Events, Trips, Place of Interest and News. These tourism terms are distributed in different trees of the schema. For example:

- Thing → Place → LocalBusiness → LodgingBusiness:
  - Hostel, Hotel, Motel, etc.
- Thing → Event:
  - MusicEvent, SocialEvent, SportsEvent, etc.

Schema.org offers other interesting terms such as customer reviews and reservation:

- Thing → Action → AssessAction → ReviewAction
- Thing → Intangible → Reservation:
  - EventReservation, FoodEstablishmentReservation, LodgingReservation, etc.

In addition, several Linked data initiatives have enabled the development of mappings between Schema.org and unavoidable semantic resources such as DBPedia, Dublin Core, FOAF, GoodRelations, SIOC and WordNet (full mappings of Web data vocabularies to Schema.org terms are available at http://schema.rdfs.org/mappings.html). We are interested in establishing the mappings at two levels. One will be a mapping between concepts which have the same name or approximate name, and between concepts which have the same role. Then taking in to account the mapped concepts (equivalentClass, subClassOf), we need to match one property from the Schema.org term with another property from a concept in TIF (equivalentProperty, subPropertyOf).

3. PROPOSED APPROACH

The tourist information can be used in different areas and different contexts. It is important that this information have to be modular, so that the elements describing the same entity can be used separately, hence the notion of a modular information represented by the IO. An *InformationObject* (IO) is a modular entity, that can be used and re-used to support tourism activities. An IO can be a hotel, a restaurant, an event, etc., or any other resource that can be attached to a tourism activity. The idea of IO is to create an entity that is:

- interoperable: can plug-and-play with any TIS,
- reusable: can be used or adapted for different use cases,
- accessible: can be stored a way that allows for easy searchability,

Table 1: Core concepts of the TIFSem ontology

CORE CONCEPTS	Description
DUBLIN CORE	The Dublin Core is a set of vocabulary terms that can be used to describe Web resources (video, images, Web pages, etc.).
UPDATE	Information about the updates of a resource, and information used to represent aspects of the evolution of the resource.
MULTIMEDIA	Information about multimedia materials associated to the described resource.
CONTACTS	Information about ways to be in relation with a resource, or a person responsible for this resource.
LEGAL INFORMATION	Information to formally identify an entity and its activities.
CLASSIFICATIONS	Information used to qualify the resource described and identify the quality of services associated.
RELATED SERVICES	Information to reference other resources.
GEOLOCATIONS	Information to locate and describe the environment a resource.
PERIODS	Information about an opening period, a closing period, a reservation period, etc.
CUSTOMERS	Customer information.
LANGUAGES	Languages spoken in the described resource.
RESERVATION MODES	Information about the reservation for the described resource, namely, who to contact customers and if the reservation is required.
PRICES	Services' price of the resource described, as well as the different means of payment.
CAPACITY	The capacity of the resource.
OFFERS SERVICES	Information about the services available within the described resource or nearby.
ADDITIONAL DESCRIPTION	Additional information about the described resource.
ITINERARIES	Activities associated to the described resource such as hiking.
SCHEDULES	Information about the availability status of services, over defined periods.

• manageable: can be tracked and updated over time.

In TIF, a typical IO is composed by 18 parts called granules (or semantic units). These granules form the main core of the TIFSem ontology. We introduce then the TIFSem ontology for the tourist domain, which contains concepts and relations for tourist resources. It consists of 19 main concepts of tourist information described by the table 1. The main concepts of that ontology have sub-concepts pertinent to the local area, and for all sub-concepts they have *is_a* relationships. Those sub-concepts have properties, restrictions and semantic relations among other concepts.

The division of these main concepts into sub-concepts, was also guided by different data sources. Indeed, in order to elaborate the TIFSem ontology, we have consulted different kinds of sources to enable the understanding and the collecting of concepts related to the specialized domain of tourism, and the corresponding vocabulary. Sources coming from Departmental Tourism Committee of the Charente-Maritime[5] (CDT17) and Departmental Tourism Committee of the Aube[6] (CDT10) were consulted.

After collecting the concepts and corresponding lexical items from the sources, we started structuring the first version of the ontology, selecting classes and subclasses (step (1), figure 4). The ontology is currently available at this link: http://flux.univ-lr.fr/public/DataTourism.owl. The mapping with Schema.org (step (2), figure 4) complements the ontology to make it more complete, up to date and coherent. Table 2 shows the mappings between the main concepts coming from TIFSem to those in Schema.org.

Table 2: TIFSem main concepts/Schemas.org

CONCEPT	Schemas.org
MULTIMEDIA	schema.org:MediaObject
CLASSIFICATIONS	schema.org:Rating
CONTACTS	schema.org:ContactPoint
LEGAL INFORMATION	schema.org:Organization
LANGUAGES	schema.org:Language
GEOLOCATION	schema.org:Place
RESERVATION MODES	schema.org:Reservation, schema.org:LodgingReservation
PRICES	schema.org:Offer, schema.org:PriceSpecification

Figures 5 and figure 6 show a simplified composition overview of the two granules *Geolocation* and *ReservationModes*, with Schema.org mapping.

The meaning of concepts is mainly expressed through the definitions in natural language. So we used OWL language which is equipped with a rich expressive power, also to enrich Schema.org which is specified in RDF. The use of an upper ontology, especially when there is extensive use of multiple inheritance, has the advantage of providing a manual navigation of the ontology and therefore simplify the recovery of concepts that can serve as a description. The last step in the development process of TIFSem is the integration of instances into the ontology coming from CDT17 and CDT10

[5] http://www.charente-maritime.org/
[6] http://www.aube-champagne.com/

(step (3) (4), figure 4).

To illustrate our approach, we consider a first example that ranks different accommodation offers based on their proximity to restaurants, bars and events in La Rochelle. A couple wants to spend a week on vacation may choose between hotels not far from downtown. The couple likes dining out, going to bars and enjoying events. Therefore, an additional characteristic that quantifies the aptness of each hotel for those that like to go out, would be of great help. However, such qualitative information is rarely available on tourism on-line Web sites. One possibility would be to compute such semantic annotations based on the geolocalisation of the hotels and those objects related to restaurants, bars and events. Figure 7 shows a simplified SPARQL query reflecting the above example. Distance function filters any potential results whose corresponding return value was not less than a constant value.

As second example, the tourist office of La Rochelle wants to have a better visibility of its territory by extracting information about rural events occurring on their territory. They need to know the events audience and information about the public attending these events. Figure 8 shows a simplified SPARQL query reflecting the second example.

4.   CONCLUSIONS

In this paper, we presented some early stage work in the Tourinflux project on identifying tourism domain ontology. Semantic technologies provide methods and concepts facilitating effective integration of tourism information originating from various sources, representing basic notions and conceptual relations in tourism domain. The objective is to improve the display of search results, making it easier for tourism industrial to find data in the right Web pages. In particular having annotations on websites that can be understood by search engines busts the on-line visibility and increases the chances that Web sites are in the search engines results to a relevant query. As part of our current and future work we are developing a Schema.org mapping to facilitate Web searching of tourism data. We are also in the process of extending the TIFSem ontology by collecting contents about more touristic service providers. We also plan to expand the TIFSem model with other types of data (about opinions, temporal data), to provide semantic and contextual answers to queries.

5.   REFERENCES

[1] T. Bittner, M. Donnelly, and S. Winter. Ontology and semantic interoperability. *S. Zlatanova and D. Prosperi (eds) Large-Scale 3D Data Integration: Challenges and Opportunities*, pages 139–160, 2005.

[2] J. D. Bruijn, D. Fensel, M. Kerrigan, U. Keller, H. Lausen, and J. Scicluna. Modeling semantic web services. *The Web Service Modeling Language*, 2008.

[3] S. Cresci, A. D'Errico, D. Gazze, A. L. Duca, A. Marchetti, and M. Tesconi. Towards a dbpedia of tourism: the case of tourpedia. *International Semantic Web Conference*, pages 129–132, 2014.

[4] M. Dell'Erba, O. Fodor, F. Ricci, and H. Werthner. Harmonise: A solution for data interoperability. *Towards The Knowledge Society: eCommerce, eBusiness, and eGovernment, The Second IFIP*
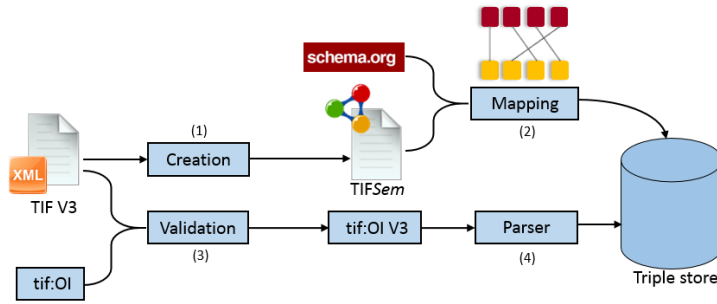
Figure 4: Simplified composition overview of the granule *Geolocation* with Schema.org mapping
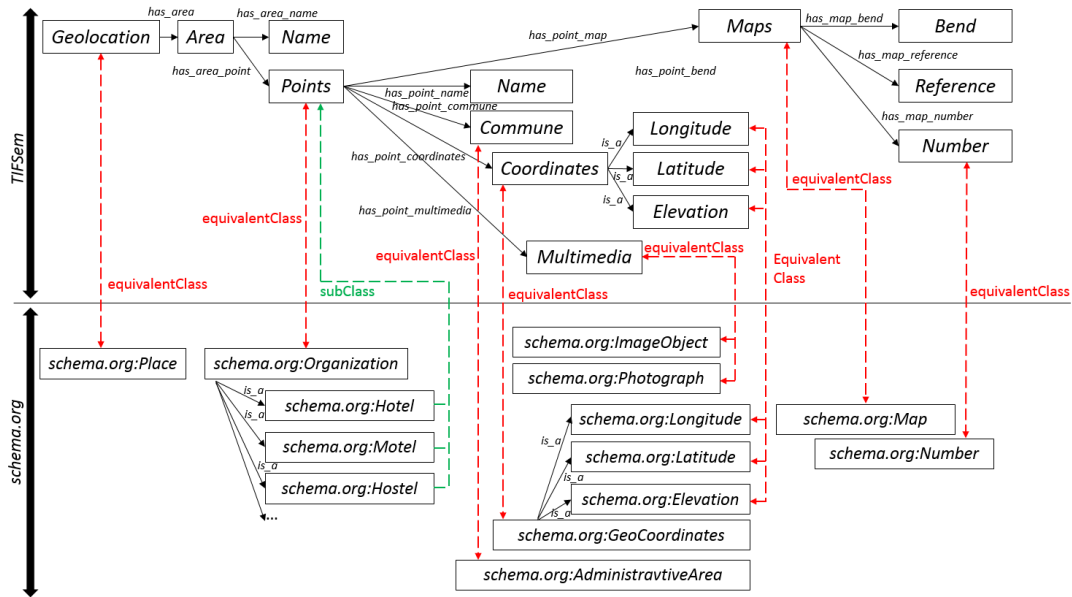

Figure 5: Simplified composition overview of the granule *Geolocation* with Schema.org mapping

Conference on E-Commerce, E-Business, E-Government, pages 433–445, 2002.

[5] G. Inkpen. Information technology for travel and tourism. *Edinburgh: Addison Wesley*, 1998.

[6] B. Legrand. Semantic web methodologies and tools for intra-european sustainable tourism. *White paper, Paris: Mondeca*, 2004.

[7] OTA. Opentravel alliance. opentravel alliance message specifications. *Specifications Document*, 1, 2000.

[8] S. Ou, V. Pekar, C. Orasan, C. Spurk, and M. Negri. Development and alignment of a domain-specific ontology for question answering. *N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias (eds) the Sixth International Language Resources and Evaluation Conference*, pages 2221–2228, 2008.

[9] I. Toma, C. Stanciu, A. Fensel, I. Stavrakantonakis, and D. Fensel. Improving the online visibility of touristic service providers by using semantic annotations. *The Semantic Web : ESWC 2014 Satellite Events, Anissaras, Crete, Greece*, pages 259–262, 2014.

[10] WTTC. World travel and tourism council, travel and tourism, economic impact 2014 world. 2014.

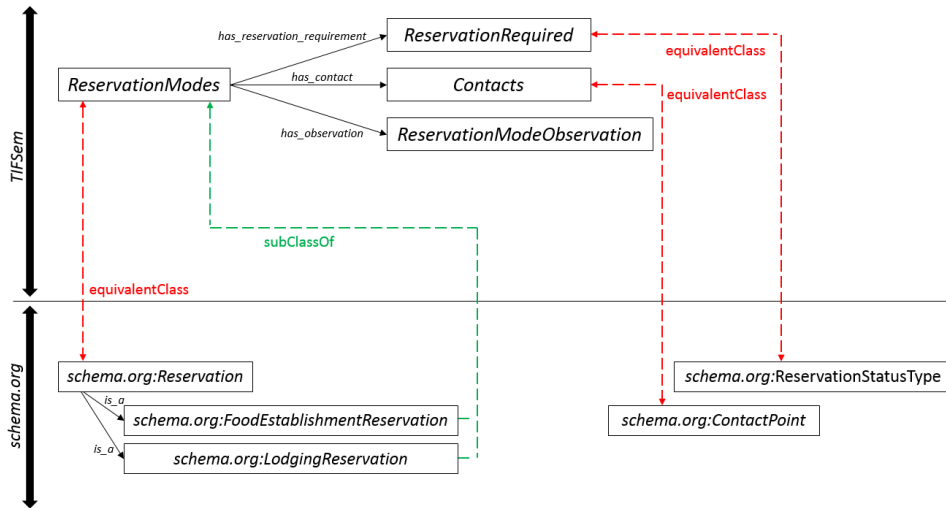**Figure 6:** Simplified composition overview of the granule *ReservationModes* with Schema.org mapping

```
PREFIX tif: <http://www.semanticweb.org/fsoual01/ontologies/2015/2/tif#>
PREFIX ldodds: <java:com.ldodds.sparql.>
SELECT ?accomodation ?restauration ?bar ?event
WHERE {
?accomodation rdf:type tif:Hotel.
?accomodation tif:hasGeolocationTownship ?township.
?township tif:geolocationTownship "La Rochelle".
?accomodation tif:hasGeolocationLongitude ?long.
?accomodation tif:hasGeolocationLatitude ?lat.
    {
    ?restauration rdf:type tif:Restaurant.
    ?restauration tif:hasGeolocationLongitude ?long1.
    ?restauration tif:hasGeolocationLatitude ?lat1.
    }FILTER (ldodds:Distance(?lat, ?long, ?lat1, ?long1) < 5).
UNION
    {
    ?bar rdf:type tif:Bar.
    ?bar tif:hasGeolocationLongitude ?long2.
    ?bar tif:hasGeolocationLatitude ?lat2.
    }FILTER (ldodds:Distance(?lat, ?long, ?lat2, ?long2) < 5).
UNION
    {
    ?event rdf:type tif:Event.
    ?event tif:hasGeolocationLongitude ?long3.
    ?event tif:hasGeolocationLatitude ?lat3.
    }FILTER (ldodds:Distance(?lat, ?long, ?lat3, ?long3) < 10).
}
```

Figure 7: Simplified SPARQL query, example 1

```
PREFIX tif: <http://www.semanticweb.org/fsoual01/ontologies/2015/2/tif#>
SELECT ?event ?audience ?maxage ?minage
WHERE {
?event rdf:type tif:Event.
?township rdf:type tif:GeolocationTownship.
?event tif:hasGeolocationTownship ?township.
?township tif:geolocationTownship "La Rochelle".
?environment rdf:type tif:GeolocalisationEnvironment.
?event tif:hasGeolocalisationEnvironment ?environment.
?environment tif:geolocalisationEnvironment "Rural"..
?audience rdf:type tif:CapacityAverage.
?event tif:hasCapacityAverage ?audience.
?maxage rdf:type tif:CustomerMaxAge.
?event tif:hasCustomerMaxAge ?maxage.
?minage rdf:type tif:CustomerMinAge.
?event tif:hasCustomerMinAge ?minage.
}
```

Figure 8: Simplified SPARQL query, example 2