Semantic Metadata for the Integration of Web-based Data for Electronic Commerce

Christof Bornhövd
DVS1, Department of Computer Science
Darmstadt University of Technology
Darmstadt, Germany, D-64283

Abstract

Today, the Internet can be seen as a global marketplace populated by a huge number of providers and consumers that exchange data from a wide range of domains. A combination of data from different sources for further automatic processing is often hindered by differences in the underlying modeling assumptions and representation. In addition, the available sources are in most cases semistructured, i.e., provide no fixed and explicitly specified schema. Therefore, an integrated use of Web-based data requires explicit information about its organization and meaning. In this paper we present a representation model well-suited for explicit description of implicitly described semistructured data, and show how this model can be used for the integration of heterogeneous data sources from the Web.

1 Introduction

Since the World Wide Web popularized its existence, the Internet has grown exponentially, leaving its roots as a researchers' forum and entering the collective consciousness. In addition to being a way for individuals and organizations to provide information, businesses have embraced the Internet as a way to offer their services. Today, the Internet is both a vehicle for advertising and a global marketplace of goods and services, ranging from electronic publications to traditional books, from financial services to travel planning, and the online monitoring of traditional logistics and physical distribution of goods.

In all these forms of electronic commerce we can identify common patterns or metaphors: In the *business-to-consumer* metaphor the business advertises and provides a service and an individual typically accesses and extracts the relevant information directly. For this kind of interaction the popular approach of presenting the information in the form of HTML pages is sufficient. The casual user browses, interprets the information and interacts with the provider in a point-and-click paradigm.

In the *business-to-business* interaction the business partners tend to rely on previously established protocols that have been in use for longer time, such as protocols for interbank fund transfers or for reservation of air travel through one of the major reservation systems, for example SABRE or Amadeus.

A third form of interaction is emerging, that may be characterized as *business-to-business-to-consumer*. A typical example of this paradigm is the search for lowest possible fares by a travel agent on behalf of a client. The travel agent is a business that acts as a knowledgeable intermediary. For this kind of service provider the typical point-and-click interaction is too time consuming while the interaction with individual reservation systems is too restrictive since many interesting opportunities are provided by ticket consolidators, last minute providers or are provided only through typical end-user oriented HTML pages. Therefore, in the business-to-intermediary-to-consumer metaphor it is necessary to extract information, consolidate it and use it for further electronic processing.

Unfortunately, the wealth of information is not provided uniformly, either because of different political and cultural contexts, or because of different intentions concerning the use of the data. The majority of data sources available online provide the information in a semistructured form, such as HTML pages. Semistructured data [1] has no obligatory and rigid schema associated with it in the sense of conventional databases. However, it provides some internal structure that is given through special tags or by the headings of sections and subsections.

An integrated use of Web-based data requires the extraction of structure and meaning of the data, the

explicit characterization of the corresponding metadata, and the consolidation of the extracted information in a common model for further electronic processing.

Our present research was motivated by concrete problems faced by the travel industry in the business-to-intermediary-to-consumer metaphor. It is our belief that this is a major growth area of electronic commerce, and that a mechanism for extracting both structure and semantics of Web-based data and making this information explicit through metadata is an essential enabler for this business model.

Most previous approaches for automatic processing of Web-based data concentrate on their structural characteristics. They are mainly based on the specification of grammars [2, 3, 10] for making the underlying structure explicit, or use browsing-oriented schemas [4, 13, 17] that represent HTML pages as objects with attributes like URL, title, and author. These approaches do not take the information content, i.e., the meaning of the data, into account in a satisfactory way.

The need for more semantic information in Web-based data has been widely recognized. Efforts, such as XML [22] try to provide a framework for additional semantic information through tags that provide hints concerning the intended meaning of the data. The use of semantic metadata for the integration of relational databases is advocated, among others, in [19, 11].

In our approach we advocate the use of existing common vocabularies or ontologies as a basis for the interpretation of Web-based data. In the travel industry these are the common three letter codes or the UNICORN protocol. Ideally, providers should adhere to those. However, in an imperfect real world, it becomes necessary to extend the existing vocabularies on the consumer side. This is quite realistic in a business-to-intermediary-to-consumer setting, since the intermediary will deal with a finite number of content providers on a regular and extensive basis. Therefore, some initial effort on the part of the consumer of the information is justified if it enables further automatic processing of the information.

In this paper we discuss the role of metadata in describing the structure and semantics of available data. We motivate our approach through a typical but simplified scenario from the travel industry. We introduce a representation model that enables the explicit description of the structure and semantics of semistructured data, and show how this model can be used for the integration of semistructured, heterogeneous Web-based data sources for further automatic processing.

2 Metadata for making structure and semantics explicit

A meaningful exchange and a correct use of Web-based data requires both information about its organization and meaning. This information, which we call context information [19, 12], provides the basis for determining the relationships between the data and the real world aspects it describes. For the explicit representation and exchange of this context information we use metadata.

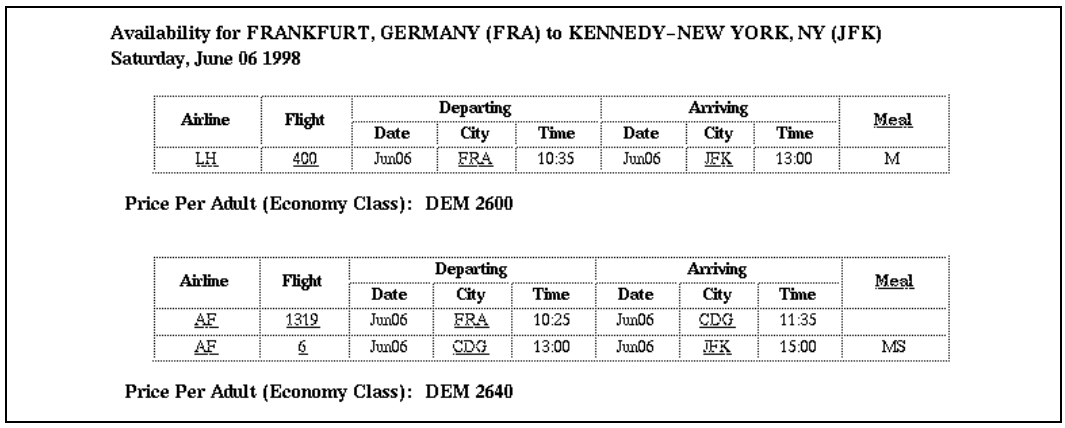
We distinguish between structural and semantic metadata. *Structural metadata* represents information that describes the organization and structure of the recorded data, e.g., information about the format, the data types used, and the syntactic relationships between them. In contrast, *semantic metadata* provides information about the meaning of the available data and their semantic relationships, e.g., data that describes the semantic content of a data value (like units of measure or scaling), or data that provides additional information about its creation (calculation algorithm or derivation formula used), data lineage (e.g., source), and quality (e.g., actuality and precision) [12].

A metadata model to describe context information in an unambiguous way is needed. We introduce domain-specific conceptualizations, or ontologies [9] that provide a commonly agreed upon vocabulary to which data providers and consumers refer. Thus, an ontology serves as a common basis for the representation of data and metadata.

Because the data we deal with is semistructured, there is no data schema available to which metadata may refer. Structure and semantics of individual data items may vary, even if they describe objects of the same class of real world phenomena. Therefore, context information concerning the organization and meaning of data has to be given on an extensional level, i.e., on the level of data values. For this reason, we need description models that allow a flexible association of metadata with the available data items. The representation model we present in Section 4 provides such a description model.

3 Scenario

In the business-to-intermediary-to-consumer scenario we are dealing with, a travel agency tries to find the lowest possible airfare by accessing information from multiple reservation systems and offerings of consolidators who represent their information differently. Figures 1 and 2 show flight information from two different online reservation systems as they are available on the Web. The available data is provided as semistructured data in the form of HTML pages.



Availability for FRANKFURT, GERMANY (FRA) to KENNEDY-NEW YORK, NY (JFK)
Saturday, June 06 1998

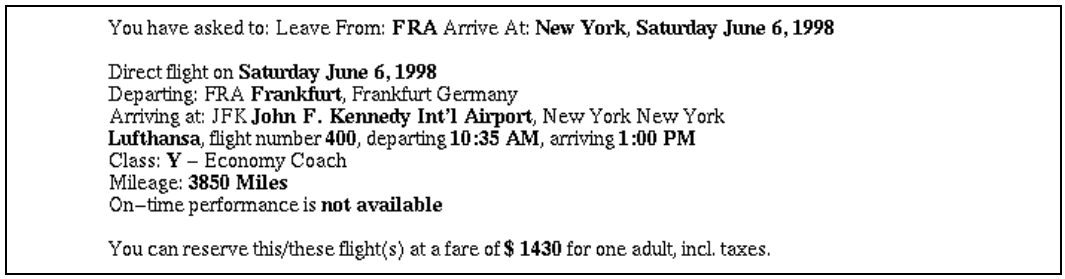
Airline	Flight	Departing			Arriving			Meal
		Date	City	Time	Date	City	Time	
LH	400	Jun06	FRA	10:35	Jun06	JFK	13:00	M

Price Per Adult (Economy Class): DEM 2600

Airline	Flight	Departing			Arriving			Meal
		Date	City	Time	Date	City	Time	
AF	1319	Jun06	FRA	10:25	Jun06	CDG	11:35	
AF	6	Jun06	CDG	13:00	Jun06	JFK	15:00	MS

Price Per Adult (Economy Class): DEM 2640

Figure 1: Reservation System A



You have asked to: Leave From: **FRA** Arrive At: **New York, Saturday June 6, 1998**

Direct flight on **Saturday June 6, 1998**
Departing: FRA **Frankfurt**, Frankfurt Germany
Arriving at: JFK **John F. Kennedy Int'l Airport**, New York New York
Lufthansa, flight number **400**, departing **10:35 AM**, arriving **1:00 PM**
Class: **Y** – Economy Coach
Mileage: **3850 Miles**
On-time performance is **not available**

You can reserve this/these flight(s) at a fare of **$ 1430** for one adult, incl. taxes.

Figure 2: Reservation System B

Because there is no obligatory data schema associated with this data, the structure underlying it is irregular, e.g., some offers are composed of multiple flight segments, and information concerning certain aspects is not given for all flights or is represented differently, as is the case with information concerning meal services in reservation system A.

Although the available data obviously has some internal structure, this structural information is not accessible as a separately specified schema, but is given in the form of HTML tags, and thus is part of the data itself. Therefore, the underlying structural information has to be extracted first to become useful for automated processing.

The data sources describe equivalent information differently. They provide different aspects of the flights, and represent the same real world aspects using different structural constructs or semantic concepts. For example, information about the flight distance is recorded in source B only, and both reservation systems identify airlines with different coding conventions. The detection and resolution of these semantic heterogeneities obviously requires knowledge about the exact semantics underlying the represented data. We were approached by a major travel agency that needed help in extracting data from the Web and in preparing it for further processing.

4 MIX — A model for explicit description of context information for semistructured data

The representation model we introduce here, called ***M****etadata based* ***I****ntegration model for data* ***X****-change*, or MIX for short, can be understood as a self-describing data model [15]. This is because information about the structure and semantics of the data is not provided as a separately specified data schema, but is given as part of the available data itself. Thus, MIX allows a flexible association of context information in the form of metadata, and is especially well suited for the representation of semistructured data.

Our model is based on the concept of a *semantic object*. A semantic object represents a data item together with its underlying *semantic context* which consists of a flexible set of meta-attributes that explicitly describe the implicit assumptions about the meaning of the data item. However, because we cannot explicitly describe all modeling assumptions the semantic context always has to be understood as a partial representation. In addition, each semantic object has a concept label associated with it that specifies the relationship between the object and the real world aspects it describes. These labels have to be taken from a com-

monly known vocabulary, or ontology. In this way, the concept label, as well as the semantic context of a semantic object help to describe the supposed meaning of the data.

The following sections introduce the fundamental concepts of the MIX model. In Section 4.1 we discuss the role of domain-specific ontologies as a common interpretation basis for data and metadata. We distinguish between simple and complex semantic objects. The concept of simple semantic objects, which are used for the representation of atomic data values, is introduced in Section 4.2. Section 4.3 deals with the idea of semantic conversion and shows how simple semantic objects can be converted among different contexts. Based on these concepts, Section 4.4 shows how conversion functions can be used for the comparison of semantic objects represented with regard to different contexts.

In Section 4.5 we introduce complex semantic objects for the representation of complex data objects. The concepts of semantic conversion and semantic equivalence are generalized for complex semantic objects in Sections 4.6 and 4.7. Finally, Section 4.8 defines the concept of semantic identity which provides the prerequisite for the integration of semantic objects that represent the same real world phenomenon.

4.1 Ontologies as a common interpretation basis

To ensure a correct interpretation of the available metadata we use domain-specific ontologies. An ontology provides an agreement about a shared conceptualization of a given application domain [9]. The concepts specified in the ontology provide a common vocabulary for which no further negotiation is necessary. In addition, the ontology provides information about the representation of the data described on the basis of the model.

In an ideal situation, all instances that make use of data and metadata from a given domain should adhere to the corresponding ontology. In an imperfect real world we must allow ontologies on consumer side that are tailored to specific needs and make the model extensible. Ontologies should use existing standards (like the UNICORN standard [20] for travel information, or the well known two letter airline code). Aspects for which no such standards exist require new ontology concepts. If a source does not adhere to existing standards or multiple standards exist, the consumer must either extend the ontology or combine existing ontologies. Depending on the application domain this can be done following a top-down approach as proposed in [7], or a bottom-up approach as introduced in [21].

In the MIX model, we simplify by understanding an ontology as a finite set of concepts and their relationships. Each ontology concept has a representation type associated with it, which is either atomic (e.g., string, integer, real, etc.), or "complex", in which case the exact representation is not determined by the concept. The domain of the representation type specifies the possible values for the representation of data corresponding to the respective ontology concept.

There is a significant difference between the terms *concept* and *type*, as they are used here. An ontology concept may be understood as an abstraction of a (homogeneous) set of real world phenomena, and thus describes the correspondence between data of a given concept and the respective domain. In contrast, the representation type determines the representation of a data value of a certain concept.

4.2 Simple semantic objects

A semantic object may be understood as a data item with additional context information attached to support its correct interpretation. For the explicit representation of context information (mainly in databases) different approaches have been discussed in the literature [18, 19, 14, 8, 16, 11]. We prefer to represent this additional information on an extensional level, because semistructured sources provide no explicitly specified data schema to which meta-information may refer.

Simple semantic objects represent atomic values, like simple number values or character strings. Based on a given ontology a simple semantic object representing value v is a 3-tuple of the form:

$$SemObj \;:=\; <C,\, v,\, \$> \;,$$

where C denotes the ontology concept to which $SemObj$ adheres, and $\$$ specifies the ***sematic context*** that records additional information which helps interpret the represented value. The semantic context is represented as a finite set of semantic objects that represent different ***semantic aspects*** which explicitly describe fundamental assumptions about the meaning and possible use of a given data object.

The following example illustrates the representation of a data value by a simple semantic object. Given an ontology that describes the meaning and representation of the concepts *Distance* (represented as a *real* value), *Unit* as the underlying unit of measure (represented as *string*), and *Scale* as the scale factor of a numerical value (also represented as a *real* value). Based on these concepts, the flight distance between Frankfurt/Main and New York as given in Figure 2 may be represented as:

$$<Distance,\, 3850,\, \{<Unit,\, \text{“}mile\text{”} >,\, <Scale, 1>\}> .$$

4.3 Semantic conversion

The association of context information with a given data value serves as an explicit specification of the implicit meaning of the data. This allows the determination of semantically equivalent semantic objects, even if they are represented differently, i.e., relative to different contexts. For example,

$$< Distance,\ 3850,\ \{< Unit,\ \text{“}mile\text{”} >, < Scale, 1 >\} > \quad \text{and}$$
$$< Distance,\ 3.85,\ \{< Unit,\ \text{“}mile\text{”} >, < Scale, 1000 >\} >$$

are semantically equivalent, because they represent the same information and we can specify a conversion function *"v [scale x] = v $\frac{x}{y}$ [scale y]"* by which one representation can be transformed into the other. Such conversion functions are a prerequisite for the integration of semantic objects coming from different sources, by converting these objects, as far as possible, to a common context.

A ***conversion function*** for simple semantic objects

$$\phi\,(\,\acute{\$},\ < C,\ v,\ \$ >\,)\ :=\ < C,\ \tilde{v},\ \tilde{\$} >$$

is a function that maps a simple semantic object, represented in context $\$$, to its corresponding representation in context $\acute{\$}$. Semantic aspects of context $\acute{\$}$ that are not specified in $\$$ are ignored for the conversion. The resulting context $\tilde{\$}$ includes the common semantic aspects plus all semantic aspects of $\$$ that are not specified in $\acute{\$}$.

For example, if ϕ_{Unit} defines a conversion function for the semantic aspect denoted by *Unit*, we get:

$$\begin{aligned}\phi_{Unit}(\,\{< Unit,\ \text{“}km\text{”} >\},\\ < Distance,\ 3850,\ \{< Unit,\ \text{“}mile\text{”} >, < Scale, 1 >\} >\,) =\\ < Distance,\ 6194.65,\ \{< Unit,\ \text{“}km\text{”} >, < Scale, 1 >\} >\,,\end{aligned}$$

with *"1 mile = 1.609 km"* being the underlying mapping rule. Conversion functions can be specified in the underlying ontology, or may be stored in an application-specific conversion library.

4.4 Semantic equivalence

The example given in Section 4.3 shows two semantic objects that intuitively appear to be semantically equivalent. However, consider the two semantic objects below:

$$< Price,\ 1430,\ \{< Currency,\ \text{“}USD\text{”} >\} > \quad \text{and}$$
$$< Price,\ 2600,\ \{< Currency,\ \text{“}DEM\text{”} >\} >\,,$$

and the conversion function $\phi_{Currency}$ that converts money according to a given exchange rate. As usual for money exchange, we have to take into consideration the asymmetry of conversion that may exist between buying and selling rates. Supposing $\phi_{Currency}$ converts US dollar to German marks on the basis of the exchange rates *"1 USD = 1.778 DEM"* and *"1 DEM = 0.55 USD"*, we get the following results:

$$\begin{aligned}\phi_{Currency}\,(\,\{< Currency,\ \text{“}DEM\text{”} >\},\\ < Price,\ 1430,\ \{< Currency,\ \text{“}USD\text{”} >\}\,) =\\ < Price,\ 2542.54,\ \{< Currency,\ \text{“}DEM\text{”} >\} >\,,\end{aligned}$$

$$\begin{aligned}\phi_{Currency}\,(\,\{< Currency,\ \text{“}USD\text{”} >\},\\ < Price,\ 2600,\ \{< Currency,\ \text{“}DEM\text{”} >\}\,) =\\ < Price,\ 1430,\ \{< Currency,\ \text{“}USD\text{”} >\} >\ .\end{aligned}$$

Because of the asymmetry of the conversion function and the difference that results when converting from *"DEM"* to *"USD"* or vice versa it may be reasonable to classify these objects semantically equivalent with regard to *"USD"*, but not semantically equivalent if currency *"DEM"* is used. This means, the result of the semantic comparison of two simple semantic objects is determined through the conversion of both objects to a common semantic context, and the comparison of the elementary data values underlying the converted objects.

Generally, the result of a semantic comparison depends on the respective semantic context used for the conversion, as well as on the conversion function to be used. We refer to the semantic context $\$$ used for the comparison as the ***target context***, and call the conversion function ϕ ***reference conversion function***, or ***reference function*** for short, of the semantic comparison.

The set of semantic aspects in the target context may be different from those in the contexts of the semantic objects to be compared. Semantic aspects specified in the original contexts but not specified in the target context are ignored for the comparison.

4.5 Complex semantic objects

Complex semantic objects can be understood as heterogeneous collections of semantic objects, each of which describes exactly one attribute of the represented real world phenomenon. These subobjects are grouped under a corresponding ontology concept. A complex semantic object that represents the complex data object o is represented as the tuple:

$$CompSemObj\ :=\ < C,\ \mathbb{A} >\,,$$

where C is the ontology concept underlying the semantic object, and $\mathbb{A}$ is the set of semantic objects associated with it that provide a representation of the subobjects of o. Again, these attributes are represented as either simple or complex semantic objects.

The attributes of a complex semantic object are divided into two distinct subsets $\underline{\mathbb{A}}$ and $\mathbb{A}_R$. $\underline{\mathbb{A}}$ is the set of key attributes that are used to identify a complex semantic object of concept C. These attributes

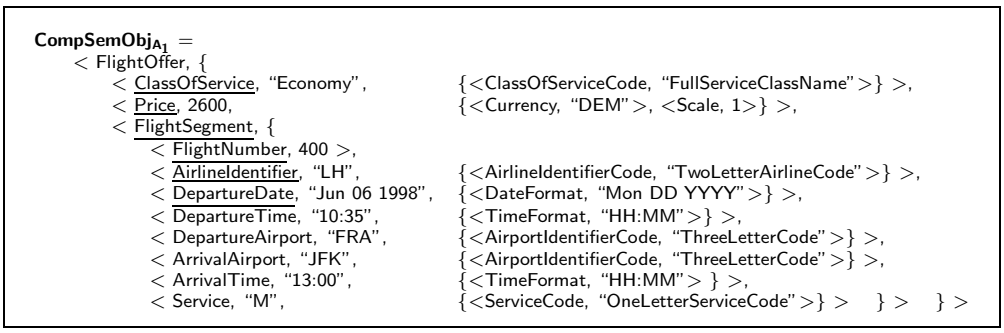


Figure 3: MIX Representation of Source A

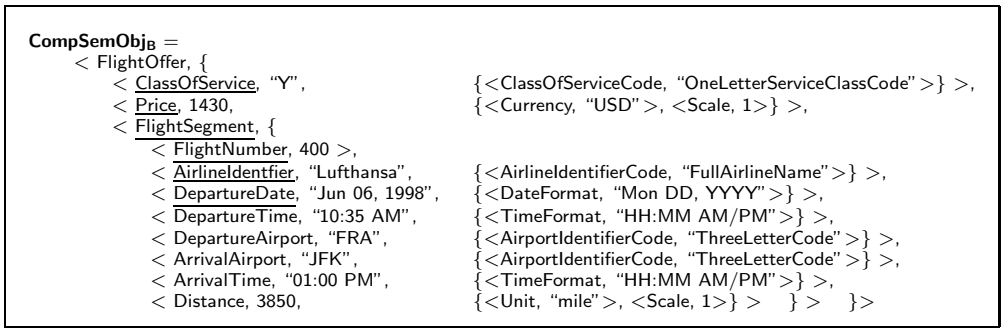


Figure 4: MIX Representation of Source B

are determined by C and specified in the underlying ontology. They provide the prerequisite for the definition of semantic identity as it is given in Section 4.8. Subset $\mathbb{A}_R$ provides the set of additional attributes. In contrast to $\underline{\mathbb{A}}$, the set of attributes $\mathbb{A}_R$ may vary between different semantic objects of the same ontology concept, as shown by the two objects above.

On the basis of an ontology the first offer given by system A in Figure 1 may be represented as shown in Figure 3 (key attributes are underlined). An offer is identified by its service class, price, and the constituting flight segments. In turn, flight segments are distinguished by their flight number, airline, and departure date. Additional properties, such as departure time, arrival airport, and meal services are not required for the unique identification of a flight segment and might not be given for all flight segments. In this way, complex semantic objects provide a flexible way to represent data with irregular structure.

4.6 Conversion of complex semantic objects

The semantic context of a complex semantic object is given through the context information specified for its subobjects. This has been defined to keep the model simple. Accordingly, the concept of a conversion function can be directly extended for the application on complex semantic objects. A ***(complex) conversion function*** Φ is a mapping function that converts a complex semantic object between different contexts by being recursively applied to all of its subobjects. If a given subobject is a simple semantic object we use the corresponding conversion for simple semantic objects.

4.7 Equivalence of complex semantic objects

In Section 4.4 we introduced the concept of semantic equivalence of two simple semantic objects. Semantic objects that are semantically equivalent represent the same information, i.e., they describe the same real world aspects.

The equivalence notion for simple semantic objects can be generalized for complex semantic objects in a straightforward manner: Two complex semantic objects with the same underlying ontology concept are said to be semantically equivalent with regard to a given target context and reference function, if their corresponding subobjects are semantically equivalent with regard to the target context and conversion function.

4.8 Semantic identity

Complex semantic objects with the same underlying ontology concept may be different because they either

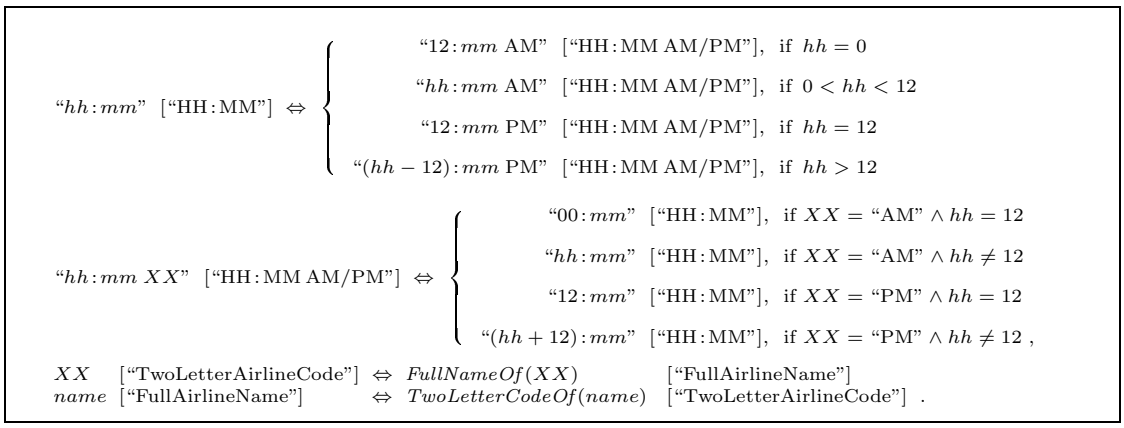


Figure 5: Mapping Rules

```
$ = {  < AirlineIdentifierCode, "TwoLetterAirlineCode" >,
       < AirportIdentifierCode, "ThreeLetterCode" >,
       < DateFormat, "Mon DD YYYY" >,
       < TimeFormat, "HH:MM AM/PM" >,
       < ClassOfServiceCode, "OneLetterClassCode" >,
       < ServiceCode, "OneLetterServiceCode" >,
       < Currency, "USD" >,
       < Unit, "mile" >,
       < Scale, 1 >                                    }
```

Figure 6: Common Representation Context

refer to different semantic contexts or because they describe different aspects of the entity they represent.

Two complex semantic objects of the same ontology concept are ***semantically identical*** with regard to a given context and a corresponding conversion function if, recursively, their *identifying* subobjects are semantically identical with regard to this context and conversion function.

At the lowest level of this recursion, simple semantic objects must be compared. Two simple semantic objects are semantically identical with respect to a given context and conversion function if they are semantically equivalent with regard to this context and conversion function, since identity and equivalence are the same for atomic values.

Thus, *semantic identity* defines whether two semantic objects describe the same real world object. In contrast, *semantic equivalence* describes whether two semantic objects represent the same information. By definition, semantically equivalent semantic objects are semantically identical since they concur in both the identifying and all other attributes. The reverse is not always true since two semantically identical objects may have the same identifying attributes, e.g., airline, flight number and date, but different non-identifying attributes, such as meal service. See Section 5 for an example.

5 Data integration on the basis of MIX

The data provided by the reservation systems introduced in Section 3 may be parsed and represented as semantic objects of concept *FlightOffer* as shown in Figures 3 and 4. By circumventing the need to agree on all attributes, the two sources will be able to agree on the same meaning for *FlightOffer*. Both data sources make different semantic assumptions (i.e. use different contexts) for the represented data.

The process of integrating data represented on the basis of MIX takes place in two steps. First, the semantic objects have to be converted to a common context, which can be specified by the application interested in the data, by using appropriate conversion functions. For example, for the aspects of *TimeFormat* and *AirlineIdentifierCode* we may specify the mapping rules depicted in Figure 5 that can be realized as functions or mapping tables.

In the second step, semantic objects which are semantically identical are identified and integrated into a common representation. Using context $ in Figure 6, as a common representation context and the conversion functions introduced so far, $CompSemObj_{A_1}$ and $CompSemObj_B$ may be classified as semantically identical because they represent the same flight offer.

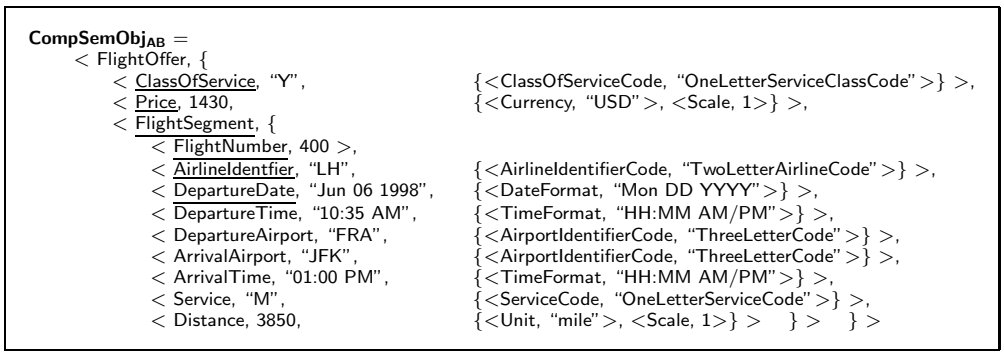


Figure 7: Unified Data Representation

Semantically identical MIX objects are interpreted as being representatives of the same real world phenomenon. Therefore, they are merged into one semantic object by unification of their attribute sets. Properties described in both objects that are equivalent are represented only once as shown in Figure 7, where $CompSemObj_{A_1}$ and $CompSemObj_B$ have been merged into $CompSemObj_{AB}$.

6 Related research

Space limitations allow us to discuss only three closely related approaches.

[19] propose a data model for the explicit representation of context information of a given data value by adding metadata that describes the organization and meaning of the data value. In addition, the model supports the conversion of this data between different contexts. The model is strictly value-based and limited to the exchange of atomic values. Thus, it lacks the possibility of defining composite objects that can be handled as one unit. Our concept of a semantic object extends the concepts discussed in [19] with regard to complex, maybe irregularly structured data objects. They assume a common vocabulary. MIX makes the common vocabulary *explicit* and provides both for the *exchange* of vocabularies, and their *extensibility*.

XML [22] provides a flexible, self-describing data model for the representation and exchange of structured and semistructured data similar to the MIX model. The XML standard supports a textual representation of data by using application-specific tags. These tags may be used to explicitly refer to the meaning of the represented data, and may be specified in a document type definition (DTD). However, XML does not enforce a semantically meaningful data exchange per se, since different providers can define different tags to represent the same or semantically similar information. Furthermore, because XML is supposed to be a very flexible though simple model for data exchange, it does not support the integration of heterogeneous data. In contrast, MIX supports an explicit representation of semantic differences underlying the data, and specifies how data based on this representation may be converted to a common representation.

In addition, MIX has some similarities with the Object Exchange Model (OEM) [15] which is a data model well-suited for the representation of data with heterogeneous structure. Besides the actual data value, each data object has a unique identifier, a type which determines its representation, and a label which provides additional information concerning the meaning associated with it. The OEM, as well as the MIX model, are self-describing data models since structure and meaning of the data objects are given as part of the available data objects. Both data models provide a highly flexible description model, especially well suited for the representation of semistructured data.

However, there are some important differences. First, in the OEM objects are identified via system-wide object identifiers. In contrast to this, data objects in MIX have certain attributes associated with them which support their identification based on their information content. Second, different from the OEM model, where data objects have source-specific labels, concept labels associated with MIX objects come from domain-specific vocabularies for which a common agreement about their meaning has been reached. These vocabularies exist and are known to users working in specific application domains. Finally, OEM is tailored mainly to the representation of data with irregular structure. In addition to this, the MIX model also supports an explicit representation of the semantics underlying the data, and provides conversion functions to convert data between different semantic contexts.

Summing up, OEM and XML provide support for the representation and exchange of data in terms of attribute/value pairs, with user defined labels. However,

this alone will not provide for semantically meaningful exchange of data, and interoperability among data providers and consumers because different providers may define their own ways of using attribute/value pairs to represent the same information. In contrast, MIX offers data providers and consumers the possibility to refer to a commonly agreed upon vocabulary, and provides hooks for the introduction of conversion functions to convert the available data to a common representation.

Unlike OEM, XML, or semantic values as introduced in [19] which can only represent object state, MIX objects include conversion functions that can be specified in the common ontology, and associated with these objects. An application may access these data objects without any further parsing.

7 Conclusion

The effective use of Web-based information by businesses requires processing beyond browsing and the common interactive point-and-click paradigm. Business users must be able to extract data for further processing. Furthermore, data from multiple heterogeneous sources must be integrated in a meaningful way by making the underlying modeling assumptions explicit.

In this paper we presented a flexible data model that supports the representation of data together with metadata that describes its organization and semantics. We showed how semistructured data can be represented and integrated by using this model. Space limitations forced us to describe a short version of MIX. A more formal presentation of the MIX model can be found in [6].

We use the MIX model in a project for integrating structured and semistructured data sources from the Internet. The prototype of a Java-based implementation exists for MIX and the MIX integration environment. Objects can be represented and integrated, displayed through a browser or used in further processing. Current research is concerned with the extension of the representation of conversion functions, and with the extraction of MIX representations for a wider range of semistructured data.

References

[1] Abiteboul, S.: *Querying Semi-Structured Data*, Proc. Int. Conf. on Database Theory, Delphi, Greece, 1997

[2] Abiteboul, S.; Cluet, S.; Christophides, V.; Milo, T.; Moerkotte, G.; Simeon, J.: *Querying Documents in Object Databases*, Journal on Digital Libraries, 1(1), April 1997

[3] Ashish, N.; Knoblock, C.: *Wrapper Generation for Semi-structured Internet Sources*, Proc. Workshop on Management of Semi-structured Data, Tucson, Arizona, 1997

[4] Atzeni, P.; Mecca, G.; Merialdo, P.: *To Weave the Web*, Proc. 23rd VLDB Conf., Athens, Greece, 1997

[5] Abiteboul, S.; Quass, D.; McHugh, J.; Widom, J.; Wiener, J. L.: *The Lorel Query Language for Semistructured Data*, Journal on Digital Libraries, 1(1), April 1997

[6] Bornhövd, C.: *MIX - A Representation Model for the Integration of Web-based Data*, Tech. Rep. DVS98-1, DVS1, Dep. CS, Darmstadt University of Technology, Germany, Nov. 1998

[7] Farquhar, A.; Fikes, R.; Rice, J.: *The Ontolingua Server: A Tool for Collaborative Ontology Construction*, Proc. 10th Knowledge Acquisition for Knowledge-Based Systems Workshop, Alberta, Canada, 1996

[8] Goh, C.; Madnick, S.; Siegel, M.: *Context Interchange: Overcoming the Challenges of Large-Scale Interoperable Database Systems in a Dynamic Environment*, Proc. 3rd Int. Conf. on Information and Knowledge Management, Gaithersburg, 1994

[9] Gruber, T.: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, International Journal of Human and Computer Studies, 43(5/6), 1995

[10] Hammer, J.; Garcia-Molina, H.; Cho, J.; Aranha, R.; Crespo, A.: *Extracting Semistructured Information from the Web*, Proc. Workshop on Management of Semi-structured Data, Tucson, Arizona, 1997

[11] Kashyap, V.; Sheth, A.: *Semantic and Schematic Similarities between Database Objects: A Context-based Approach*, VLDB Journal, 5(4), 1996

[12] Madnick, S.E.: *From VLDB to VMLDB (Very MANY Large Data Bases): Dealing with Large-Scale Semantic Heterogeneity*, Proc. 21st VLDB Conf., Zürich, Swizerland, 1995

[13] Mendelzon, A. O.; Mihaila, G. A.; Milo, T.: *Querying the World Wide Web*, International Journal on Digital Libraries, 1, 1997

[14] Ouksel, A.M.; Naiman, C.F.: *Coordinating Context Building in Heterogeneous Information Systems*, Journal of Intelligent Information Systems, 3(2), 1994

[15] Papakonstantinou, Y.; Garcia-Molina, H.; Widom, J.: *Object Exchange Across Heterogeneous Information Sources*, Proc. Int. Conf. on Data Engineering, Taipei, Taiwan, 1995

[16] Rosenthal, A.; Sciore, E.: *Description, Conversion, and Planning for Semantic Interoperability*, Proc. IFIP WG 2.6 Working Conference on Database Applications Semantics (DS-6), Atlanta, Georgia, USA, 1995

[17] Smith, D.; Lopez, M.: *Information Extraction for Semi-Structured Documents*, Proc. Workshop on Management of Semi-structured Data, Tucson, Arizona, 1997

[18] Siegel, M.; Madnick, S.: *A Metadata Approach to Resolving Semantic Conflicts*, Proc. 17th VLDB Conf., Barcelona, Spain, 1991

[19] Sciore, E.; Siegel, M.; Rosenthal, A.: *Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems*, ACM TODS, 19(2), June 1994

[20] UNICORN Maintenance Authority: *UNICORN Application Standard (TTIP03) V4.0*, Travel Technology Initiative Ltd., c/o Cosmos Management Services Department, Bromley Kent, 1994

[21] van der Vet, P.E.; Mars, N.J.I.: *Bottom-Up Construction of Ontologies*, IEEE Trans. on Knowledge and Data Engineering, 10(4), 1998

[22] World Wide Web Consortium: *Extensible Markup Language (XML) 1.0*, Feb. 10, 1998