Metadata-oriented Data Model Supporting Railway Distributed System Integration

Hanning Wang
State Key Laboratory of Rail Traffic Control and Safety, Beijing JiaoTong University
Beijing, 100044, P.R.China
Email: 06114158@bjtu.edu.cn

Weixiang Xu and Chaolong Jia
School of Traffic and Transportation, Beijing JiaoTong University
Beijing, 100044, P.R.China
Email: {wxxu, 07114214}@ bjtu.edu.cn

***Abstract*—Railway distributed system integration needs to realize information exchange, resources sharing and coordination process across fields, departments and application systems. And railway data integration is essential to implement this integration. In order to resolve the problem of heterogeneity of data models among data sources of different railway operation systems, this paper has brought forth the concept of metadata dictionary that abstracts related information and information content of each data source. Based on the metadata dictionary, this paper presents a novel integration data model of spatial structure, a XML-oriented 3-dimension common data model. The proposed model accommodates both the flexibility of level relationship and syntax expression in data integration. In this model, a spatial data pattern is used to describe and express the characteristic relationship of data items among all types of data. Based on the data model with rooted directed graph and the organization of level as well as the flexibility of the expression, the model can represent the mapping between different data models, including relationship model and object-oriented model. A consistent concept and algebraic description of the data set is given to function as the metadata in data integration, so that the algebraic manipulation of data integration is standardized to support the data integration of distributed system.**

***Index Terms*—distributed System, integration, data model, metadata, data Integration.**

I. Introduction

Railway distributed system integration is dedicated to integrating discrete information resources, realizing information exchange, resources sharing and collaborative processes across fields, departments, application systems, and providing powerful support to railway transporting organizational optimization design and operation management. Over several years of information construction, Chinese railway system has established MIS (Management Information System) in various business departments, such as the FTMS (Freight Transportation Management System), PTMS (Passenger Transportation Management System), TDCS (Train Dispatching Command System), ATIS (Automatic Train Identification System), LMIS (Locomotive Management Information System), CSMIS (Communication and Signal Management Information System), PWMIS (Permanent Way Management Information System) etc.[1]. Usually these MIS of various business departments (especially legacy systems) deal with particular business problems respectively, which leads to the fact that each adopted architecture, implementation language, integration point and interaction protocol are different from each other, resulting in the complication for distributed application integration, as shown in Fig. 1.

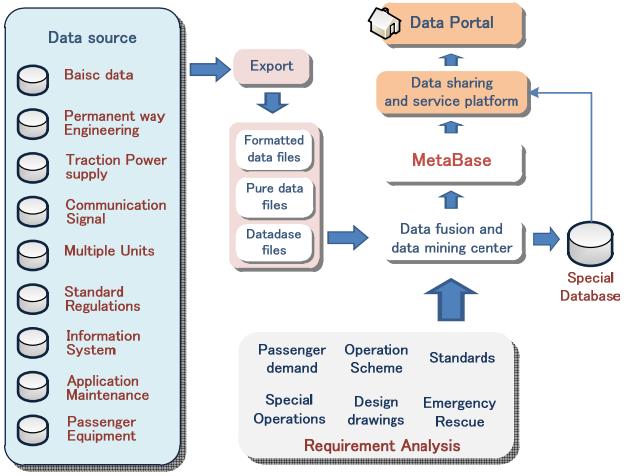

Figure 1. Multi-source railway data integration.

In order to solve this problem, we need to break the problem down into several layers, and the integration problem in each different layer follows a different integration approach. Based on different integration points, those layers are usually divided into Transmission Mechanism, Data integration, Interface Integration and Process Integration from bottom to top [2, 3, 4]. Data transmission and data conversion are basic functions of the Data Integration layer, where unified criteria and specification are emphasized in classification, processing and representation of various data information in order to realize a seamlessly linked information interface. From

doi:10.4304/jsw.7.4.814-822

the perspective of layers for various business system integrations, interface integration and process integration are built on the basis of railway data integration. Thus the core problem in railway distributed system integration is to figure out how to describe the data structure, release and lookup data information; how to consistently represent and standardize the operation of data information of heterogeneous platforms, different formats and semantics; how to scientifically integrate large amount of data information in the network environment per user requirement, to realize the barrier-free interactive collaboration of data information.

II. STATUS QUO AND FEATURES OF THE RAILWAY DATA INTEGRATION STUDIES

In terms of rail traffic data mergence and integration key technology study, Nash A, et al. have studied a data exchange markup language RailML for different railway application systems. It is developed based on XML language and simplifies the data transmission process by making use of the common data structure [5]. The Japanese scholar Suzuki Satoshi has studied the vectorization of the rail traffic space data, and produced a boundary tracing algorithm and a linear tracking algorithm [6]. For the large-scale map, based on related data standard, it's capable of digital vectorization by layer and distinguishing the railway line and other geographic objects. The American scholar Harrison K. et al. have studied the comprehensive utilization of digital elevation model (DEM), satellite remote sensing data, aerial photographs data and other geospatial data to represent the topography and land coverage and establish visualized 3-D environment for rail transportation [7]; Dell' Aquila et al. from Italia have constructed a data warehouse for decision support based on the national railway transportation data [8]. They have studied data acquisition, data management and related issues for the data warehouse, and analyzed the problem of relation-type query results data scale in the data warehouse, with a back-end and front-end solution solving the problem of query efficiency. Brian L. Smith et al. from University of Virginia, America have studied the approach of data extraction, transformation and loading (ETL in short) of the transportation data warehouse [9], and discussed the four basic components of the ETL process: data aggregation, data quality assessment, and data source tracking and data characterization, given their realization method.

Railway data includes two major types of data, i.e. static and dynamic. The dynamic data mainly includes the railway permanent way equipment test and maintenance data in a fairly long cycle (5 or 10 years), passenger ticket data and freight bill data in transportation department. While the static data includes railway basic geographical information data, remote sensing image data, station site and line equipment comprehensive data, as well as various topology relationship data (line network topology, track and equipment point topology relationship, and insulation section and track circuit topology relationship etc.) from all departments of each railway bureau all around the country. Data is coming from various railway departments, which shows features such as huge data amount, scattered storage, different types and various formats. For example, there are relational database files such as Oracle, SQL Server, Sybase, Access, document files, waveform graph files with an extension of GEO generated by the track inspection cars in works department. Also there are spatial data files such as AutoCAD, visio, MapInfo and Arc/GIS, graphic files, route video files and remote sensing image files etc.. Files generated by these different systems all have different approaches for definition, expression and storage of their object. The discrepancy among various types of data files has caused great inconvenience for information sharing. Thus the solution to transformation among multiple data formats has always been the important pending problem for data platform system development in recent years.

From the macroscopic perspective, data sharing means that consistency could be achieved in terms of data format, syntax and semantics. For the purpose of supporting multiple application system, various data concepts have to be merged together to establish a unified integrated data model, which can:

- Describe the abstract structure, specific expression and relation of data formats, support various data structure models with different semantics and formats, and solve the data diversity in application.
- Provide the metadata description for both atom data and combination data. In distributed system integration, the integrated data model should be as simple as possible to make the conversion of data models among various data sources more convenient. The integrated data language based on common data model should conveniently represent the source data and its transaction process, be able to support and process set-oriented statements.

III. THE STRUCTURE OF A METADATA DICTIONARY

Metadata is data for describing data [10], which is referred within this paper as the related information on describing database location, identification of data, coverage scope, quality, spatial and temporal patterns, provision method of data etc. Responding to user's transparent integration access demand in respect of data location, data identification and data model, integrated system data access interface should be able to provide users with abstract public logic entity that can be queried and can eliminate the property description conflict in the data model. When the user launches a query request based on the property of the logic entity, the integrated system can decompose and map this query for integrated environment into a specific data source. This process is dependent on the right expression of related information for every data source under the integrated environment. Therefore, this paper constructs a metadata dictionary under the distributed heterogeneous data integration service framework to describe/express, store and access these non-data information. Meanwhile, a metadata access engine is provided, giving necessary support to

apply and retrieve metadata, decompose query and construct query sentences.

Setting aside the different external technical representation and discrepancy in data structures for railway's each component system data source, the information contents recorded by each data source under the integration environment does have relevance. Based on scattered information content, the metadata dictionary abstracts multiple entity subsets from each data source and defines their properties that are feasible and should be owned. And then a visualized and straightforward naming mechanism is applied to name those entity subsets as the logic entity sets that can be queried for users. The railway metadata logic entity is an aggregation entity (MD_Metadata),mainly including six entity subsets components which are identification information (MD_Identification), data quality information (DQ_DataQuality), spatial reference system information (RS_ReferenceSystem), content description information (MD_ContentDescription), distribution information (MD_Distribution), responsible party contact information (RP_ResponsibleParty) and physical description information (MD_PhysicalDescription). Their logic structure is shown in UML class diagram as follows.

Since the railway data includes both spatial information and non-spatial information, while the description of non-spatial information involves no spatial reference system information, the corresponding entity can be selected. As shown in Fig.2, the MD_Metadata aggregation entity itself contains three required elements, i.e. the date, the contact information and the mapping information MapSchema. The MapSchema element, which is used to indicate the mapping relationship between logic property and physical property, is a complicated element. It determines the physical property that is mapped to by a logic property with the path of storage equipment, data source, relationship table and physical fields, including the storage equipment, database, table and physical fields where the physical property belongs. Thus, it can provide support for decomposing and mapping query process for integration environment. MD_PhysicalDescription keeps a detailed record of each physical database's related information, including network location, access interface, data model etc., with a specific level down to relationship table and field. Besides the IP element indicating the database network location, the Provider element for database programming access interface and the table element describing the relationship table inside the database, the relation element is also used to describe the associated relationship between each two tables. For each relationship table, the relation element's "to" property indicates the external key of the table, "ref" property indicates the table name associated to the external key. The primary key defined by the table can be found through the table name in order to establish the reference relationship.

The physical entity is the data object tightly associated to the underlying data source. But the mapping relationship between logic entity property and physical property, which is established through the MapSchema element of the logic entity, has shielded the underlying data's features such as organization structure and naming mechanism from higher layers. Furthermore, when updating the database schema, it is ensured that only this mapping mode needs to be updated. It makes sure the relative independency of the physical entity and can meet the demands of future technology development, database content updates and continuous development.

IV. DATA MODEL BASED ON A DIRECTED GRAPH

In a distributed system integration, in order to deal with the heterogeneous nature among component system data models, the most widely used method is to provide an effective Common Data Model（CDM） so as to describe the global schema and global query language. A fairly great part of the existing railway business data is

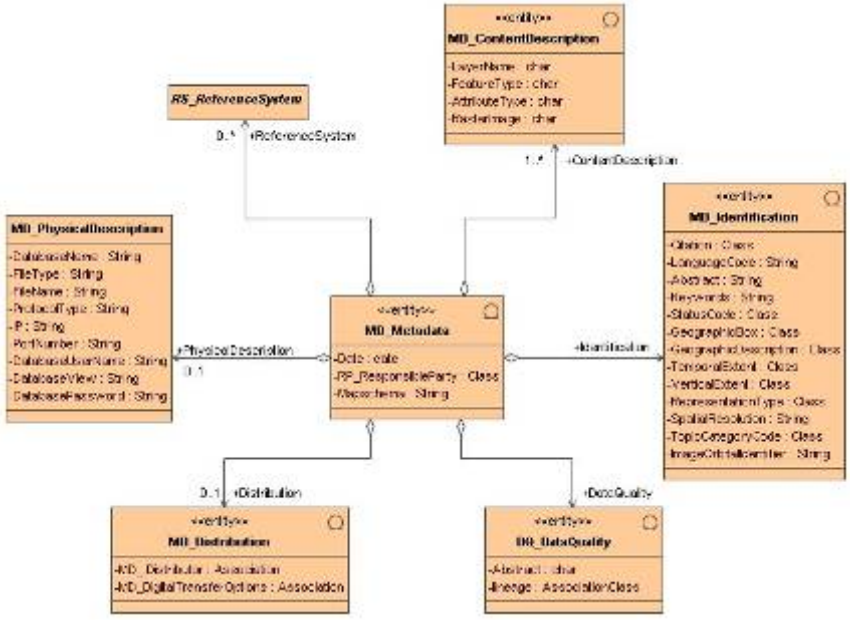

Figure 2. Metadata UML class diagram

semi-structured or non-structured data, therefore traditional data integration technologies are proved unable to function as expected. For example, multiple database CORDS [11] adopts the relationship model, but not all the source data can be expressed through relationship. MIND [11] adopts the traditional Object Oriented (OO) model, which owns stronger description ability than the relationship model, but lacks a good self-describing ability. What the OO data model can manage is mainly for the structured data, while the data stored in HTML/XML documents and multimedia is more difficult to describe with the OO data model. Besides, the data pattern and data are separately stored in OO model, which makes self-describing data integration inconvenient.

Currently most researchers tend to adopt the data model based on the directed graph, where the relationship between data is described with node and edge in the directed graph. Different from the traditional data model, data description information and data are stored together in this type of data model, so that it's more suitable to be used to describe those data without explicit structure.

TSIMMIS[12] is a well-known heterogeneous information source integration system researched and developed by the Database Research Team of Stanford University. In the OEM object model, an object is described with a quadruplet $\langle OID, label, type, value \rangle$ .Where: *OID* is an object identifier, $label$ is used to describe the meanings represented by the object; $type$ is used to describe the type of the object; $value$ is used to describe the value of the object. In order to be able to describe the data from various types of data sources, $type$ in an object descriptor $\langle OID, label, type, value \rangle$, which is able to represent basic data type, can also represent the data set type (e.g. set, list) and reference type. If the type of an object is reference type, which implies that the object is congregated by other objects, its value is the set of identifiers referenced by it.

Through a root connected directed graph, an OEM object (O) can be represented as $O(r, V, E)$ .Where the vertex represents the object; the edge represents the reference relationship between objects. The root node is an aggregate object that references types; V is the set of the aggregating object and its all referenced objects; E is the set of reference relationship between objects, that is to say,

$$E = \left\{ \begin{array}{l} \langle v_i, v_j \rangle \mid v_i \text{ is the identifier of object } O_i \text{ in } V \\ \wedge v_j \text{ is the identifier of object } O_j \text{ in } V \wedge \text{the} \\ \text{referenced object } O_j \text{ of } O_i \text{ is the identifier of} \\ \text{object } O_j \text{ in } V \end{array} \right\}$$

The XIDM model produced by Huazhong University of Science and Technology in the Panorama project is an integrated data model based on XML. The global model of an integrated system can be represented by a marked directed connected graph $G = \langle vertex, edge \rangle$ in XIDM. Where: $vertex$ is the set of nodes; $edge$ is the set of edges.

A node of graph G can be represented by a 5-tuple array

$$ecluster = \left\langle \begin{array}{l} key, attributes, subEClusters, \\ qualifications, mappings \end{array} \right\rangle .$$

Where: key is the list of key attributes of element cluster; $attributes$ is the ordered list of attributes of the element cluster; $subEClusters$ is the ordered list of $ecluster$'s sub-element cluster; $qualifications$ is the set of restriction conditions met by $ecluster$'s elements; $mappings$ is the set of $ecluster$'s model mapping information.

The OIM[12] model produced by Southeast University in the *Versatile* project is also a data model based on directed graph. OIM object algebra is produced on the basis of defining OIM object model. OIM object algebra has defined a series of operations for OIM object, including the Union, Difference, Selection, Projection, Join and Division.

Additionally there are several data models that make improvement and development for the abovementioned models. Although the study for these data models have made a great progress and got fairly wide application, they still have some shortages with regard to the implementation of railway distributed system integration to support various data with different semantics and formats. And those shortages are mainly shown in the following aspects: the semantics relation of the object identifier have not been taken into consideration; no complete algebra operation definition; no definite representation order for the description of objects in the same layer; insufficient support for the object representing attribute-value pair; lack of unification for the entire structure specification required by the data integration; incapable of consistent data representation and operation in terms of data format, data syntax and semantics; insufficient capability of data scalability. This paper combines both the advantages of OEM model and OIM model, utilizing the spatial data mode to describe and represent the relationship characteristics among various types of data. The XML-oriented 3-dimension common data model (X-3DCDM) is established to function as the role of metadata in data integration.

V. X-3DCDM MODEL

A. Description of the object in X-3DCDM model.

In order to solve the problem of heterogeneous railway information data integration, an XML oriented three dimensional common data model (X-3DCDM) is designed in this paper. Data objects in this data model can be grouped into 4 types based on the space dimension, which are respectively:

$X-3DCDM$: α | β | γ | ε

α : 0-dimensional (0D) space object set, α : 0-dimension (0D) space object element, namely basic data type (Atom), which represents the *point* object, mainly including NUMBER, REAL, INEGER, STRING, BOOLEAN, BINARY etc.

β : 1-dimensional (1D) space object set, β : 1-dimension (1D) space object element, namely simple data type (Simple), which can be the object of *Line*, *String* or *Points*; It is defined based on combination of basic data types; Each key word has its own different rule based on up or down mortals respectively. All these rules conceal in data semantic restraints in space construction. It can be expressed in semantics restraint explicit scheme, with the uniformity in data, so the exchange semantic can be guaranteed.

γ : 2-dimensional (2D) space object set, ν : 2-dimension (2D) space object element, namely integrated data type (Integrate), which can be the object of *region* or regions; It's defined based on combination of simple data types with corresponding key word, and it's a linear set of simple data types.

ε : 3-dimensional (3D) space object set, ε : 3-dimension (3D) space object element, namely the construction data type (Constructing), which can be Point, Line, String, Constructing; The construction data type can be constructed by an atom data type, a simple data type, or an integrate data type that has specific meta-data semantic. As shown in Fig.3.

The relationship between these 4 types of data object is as follows:

$\beta \subseteq \alpha = \{\alpha_1, \alpha_2, \alpha_3, ...\};$

$\nu \subseteq \beta = \{\alpha_1, \alpha_2, \alpha_3, ...\};$

$\varepsilon \subseteq (\gamma \vee \beta \vee \alpha) = \{\{\nu_1, \nu_2, ...\} \vee \{\beta_1, \beta_2, ...\} \vee \{\alpha_1, \alpha_2, ...\}\}$

The data object types of X-3DCDM are grouped into basic data type, simple data type, integrated data type and construction data type. Object descriptor varies with the different types that the object belongs to, with specific semantic description as follows:

$X-3DCDM$:*Atom* | *Simple* | *Integrate* | *Constructing*

$Atom: \langle tag, oid, type, value \rangle$

$Simple: \langle tag, oid, type, a-oid-list \rangle$

$Integrate: \langle tag, oid, type, s-oid-list \rangle$

$Constructing: \langle tag, oid, linktype, ref-oid-list \rangle$

Where oid is the unique object identifier; The XML identification tag is similar to the $label$ in OEM model, indicating the meanings that the object represents; $type$ is the object type, and for the Atom object, it can denote the original data type and basic type allowed by XML Schema (excluding types such as ID, IDREF / IDREFS, ENTITY / ENTITIES, NOTATION). For complicated object, $type$ can denote the set data type; $Value$ is the value of basic type, and the complicated object value $a-oid-list$, $s-oid-list$ and $ref-oid-list$ is the list of a group of $\langle rank, oid \rangle$, representing the sub-object set included by this object, where $a-oid-list$ is $rank::0|1$, with the value of 0 of $rank$ indicating that the sub-object is the basic type, and the value of 1 of $rank$ indicating series of basic type; and $s-oid-list$ is $rank::0|1|2$, with the value of 0 of $rank$ indicating that the sub-object is the basic type, and the value of 1 of $rank$ indicating the series of basic type, and the value of 2 of $rank$ indicating the linear series of simple type. And $ref-oid-list$ is $rank::0|1|2|3|4$, with the value of 0 of $rank$ indicating that the sub-object is basic type, the value of 1 of $rank$ indicating series of basic type, the value of 2 of $rank$ indicating the linear series of simple type, the value of 3 of $rank$ indicating series of aggregated data type, and the value of 4 of $rank$ indicating series combination of the simple type (basic type) and aggregated data type.

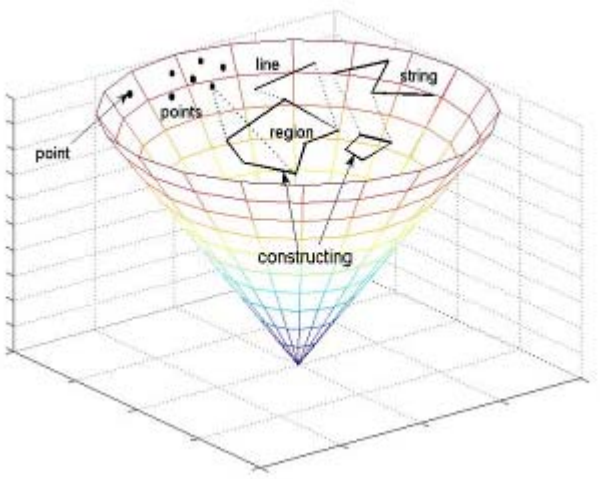

Figure 3. Schematic diagram of Data objects

Based on the types of $Constructing$ links, the types of reference object can be divided into internal reference (corresponding to the IDREF, ID, FS attribute of XML element or keyref restraint defined by XML Schema) and external link (corresponding to XLink), which include simple link and extended link. Therefore, three types of reference are defined in X-3DCDM: $inner$, $simple$ and $extended$, respectively representing the internal reference, simple link and extended link. For the reference object with a link type of $inner$, its value is the oid of the reference object; For the object with a link type of $simple$ and $extended$, the object value is similar to the value of a complicated object, namely a list of a group of oid , which denotes the sub-element,

attributes and character data included by this reference object.

B. The object algebra of X-3DCDM

The formalized definition of X-3DCDM algebra object description is as follows:

$$X = \left((V, V_{root}), E, A, Rule, Ref, EAstr\right)$$

Where:

$V = \{v_1, v_2, ...v_n\}$ is a V_{root} finite element set, of which the $v_i (1 \le i \le n)$ is the element in the XML data set;

$E = \{e_1, e_2, ..., e_n\}$ is the set of edges, of which the $e_i (1 \le i \le n)$ is the element in the set of data relationship edges;

$A = \{a_1, a_2, ..., a_n\}$ is the set of attributes, of which the $a_i (1 \le i \le n)$ is the attribute in the XML data set;

$Rule(V)$ is the set of XML data set rules;

$Ref(V_1/A_1, V_2/A_2)$ is the quotation relationship on XML data set;

$VAstr$ is a mapping from V to the power set of A.

All of the XML documents compiled based on the X-3DCDM algebra object description model are the specific instances under this model - an ordered XML document.

The formalized definition of an instance is shown below: Assuming that NV is the set of finite element points in an XML document, three triple (NV, NV, NV) can be used to record the order relation between fathers and sons as well as brother and brother among element nodes. For $(n_p, n_l, n) \in \{(NV, NV, NV)\}$, it means that for the node n, its father node is n_p, left brothers is n_l; If n doesn't have any left brothers, $n_p = n_l$. $[n_1, n_2, ..., n_k]$ is a sequence of all sons belonging to n_p ordered from the left to right, which is denoted as $Extend(n_p)$. $Extend(n_p) \in \{(NV, NV, NV)\}^*$, which is a sequence that could be inferred from the regular expression $Rule(n_p)$, and this is recorded as $Extend(n_p) \Leftarrow Rule(n_p)$ [13].

The example of the X-3DCDM data set is denoted as: $XS = (NV, E_{NV}, NA, Ref_{NV}, NEAstr, root)$

Where:

NV is a set of the finite element nodes, enabling such a mapping to exist: $tag : NV \to E$;

NA is a set of the finite attribute nodes, enabling such a mapping to exist: $att : NA \to A$;

$E_{NV} \subseteq \{(NV, NV, NV)\}$, for any $Extend(n_p) \in E_{NV}^*$, it always exists that $Extend(n_p) \Leftarrow Rule(n_p)$;

$Ref_{NV} \subseteq \{(NV / NA, NV / NA)\}$, given that $\forall (nv_1 / na_1, nv_2 / na_2) \in Ref_{NV}$, it always exists that $(tag(nv_1) / att(na_1)), tag(nv_2) / att(na_2)) \in Ref$;

$NVAstr \subseteq \{(NV, NA)\}$, given that $\forall (n_1, n_2) \in NVAstr$, it always exists that $att(n_2) \in VAstr(tag(n_1))$;

Given that $root \in NV$, it exists that $tag(root) \in V_{root}$.

X-3DCDM model supports 6 sorts of object operation provided by OIM model, and also supports 2 sorts of algebra operations of object *Intersection*, object *Division* for expand OEM. See the reference [13]. Data integration is completed by integrating a series of different types of (or heterogeneous) data sets, therefore, operations on data object would be taken as algebra operations on different data sets, which forms the rules of integration. For the given $XS_1 = (NV_1, E_{NV1}, NA_1, Ref_{NV1}, NVAstr_1, root_1)$, $XS_2 = (NV_2, E_{NV2}, NA_2, Ref_{NV2}, NVAstr_2, root_2)$,…, $XS_n = (NV_n, E_{NVn}, NA_n, Ref_{NVn}, NVAstr_n, root_n)$ they are homogeneous or heterogeneous data with a total number of n. According to the $Rule$ set, there is: $X = (NV, E_{NV}, NA, Ref_{NV}, NVAstr, root)$, and this process is the reconstruction of data.

X-3DCDM provides the data expression and operation specification in a unified standard for the data integration in the distributed heterogeneous system, setting forth the formal description of specific syntax, semantics and algebra operation, and having achieved consistency in three aspects of data format, syntax and semantics. Therefore, X-3DCDM owns greater flexibility and is appropriate for the integration of heterogeneous systems, which is reflected by the following facts:

• X-3DCDM has described the abstract structure, specific representation and relation of the model, summarizing, abstracting and extracting 4 sorts of data object types, which can embrace multiple types of heterogeneous data structure and complicated data types, and thus it solves the problem of railway business data diversity.

• X-3DCDM has defined the semantic relation and syntax expression of the model, which can generate consistent data representation standard and data operation convention between different fields, different businesses and different departments in the process of railway information construction, and by using the metadata it has built a consistent data model for heterogeneous system integration in order to deal with the difficulty of data sharing and system interaction between existing systems arising from the lack of consistent data structure convention.

• X-3DCDM details multiple algebra operations of the model object, including *Union*, *Difference*, *Intersection* etc., which has realized the reconstruction of data objects, providing a formal basis for the query decomposition and query optimization of the data object. Whether the two objects are compatible with each other or not, operations such as object *Union*, *Difference* and *Intersection* can now be performed on them, realizing the mapping to data in different granularity and different model, such as the mapping between relation model and object-oriented model, and the connection relationship of one-on-one mapping between heterogeneous systems is established. This guarantees that the heterogeneous system data information can be reorganized based on the syntax, semantics and algebra operations provided by X-3DCDM, ensuring an accurate complete "reconstruction", thus it can effectively support the comprehensive integration of heterogeneous systems.

VI. X-3DCDM- BASED INSTANCE ANALYSIS

High-speed railway metadata is the standardized description of all high-speed information resources. When metadata content are needed to be exchanged or processed between homogeneous or heterogeneous systems in the form of automation, the metadata for high-speed railway data sharing is required to be defined with the XML model. This is for the convenience of storing, exchanging and sharing the data between various business departments in a consistent manner.

In Section 3, the high-speed railway metadata has already been abstracted with UML and a metadata UML model is formed, which model is irrespective of and neutral from the syntax. On the basis of this model and according to the syntax specifications of W3C XML model, the metadata entity, metadata type entity, metadata element and code table in the metadata of the metadata's UML model are mapped into an XML model. And through the X-3DCDM model proposed in the paper, the XML could be further transformed into an orderly directed tag map.

A. The detailed matching method is described as follows:

(1) An XML element corresponds to an object in X-3DCDM model. Thus the nesting relationship between elements in XML would reflect the relationship between objects and sub-objects within the model.

(2) An XML element *tag* corresponds to the tag of the descriptor for X-3DCDM model object, indicating the implication represented by the object.

(3) XML element's attribute name-value pairs (excluding the attribute value pairs of the reference type) correspond to the atomic object in X-3DCDM model. The attribute name is used as the object identifier.

(4) The XML link references are divided into internal reference and external link (including simple link and extended link), which correspond to the reference objects in X-3DCDM model.

(5) As for hybrid content element (i.e. a hybrid containing both character data and sub-element) in the XML document, it can be treated as a complicated object, where the character data is treated as an individual atomic object parallel to other sub-elements, with TEXT being its identifier and the character data as the value of such atomic object.

(6) Since XML elements are orderly, it's defined in the X-3DCDM model that all object nodes are to be arranged in accordance with such a strategy that is from left to right where the span degree overrides.

(7) The sub-element type and element attribute contained in an XML element could be reflected by the *rank* value of a complicated object in the XCDM model. The *rank* value as the edge attribute could be placed altogether with the object *tag* on the edge of a directed graph. Through the *rank* value, X-3DCDM model can provide the user or applications with multi-tier views, e.g. tree structure at the character level and chart structure at the semantics level etc.

B. Key points in describing an X-3DCDM model's directed graph:

(1) The node of an X-3DCDM model's directed graph is required to has a unique object identifier *oid* , which can correspond to the *id* attribute of an XML element. For the X-3DCDM model's directed graph's edge tag graph, the element *tag* in an XML document (i.e. tag in an object descriptor) is labeled on the directed graph's edge.

(2) There are two types of edges in an X-3DCDM model's directed graph, namely edges of *tag* type ($rank :: 0 \mid 1$, presented as a solid edge) and edges of the reference type ($rank :: 2 \mid 3 \mid 4$, presented as a broken edge).

(3) In the XCDM model's directed graph, a node of the atomic object and complicated object is represented with a solid circle, and object of the reference type is represented with a broken circle (excluding the internal reference objects).Edges of the reference type and solid circle comprise the internal reference relationship between complicated objects.

(4) Value of the Rank is labeled on the X-3DCDM model's directed graph's edge as its property, with the default value being 0 that's could be neglected in the graph.

Based on the matching method above, an X-3DCDM model directed graph can be obtained for the following paragraph of XML document for distributor and contact in metadata distribution information. See figure 4.

```
<distributioninfo>
<distributor id='2'contact='3'>
<!- Internal IDREF reference, and rank has a value of 2- >
<individualname>Ming Zhang</individualname>
< MD_Source xlink: type=" simple" xlink;show="replace"
xlink：href=homepage.xml" xlink：title="MD_Source"/>
< !-Simple link, where the link type is simple, and rank has a value
of 3.->
<email>MingZhang@bjtu.edu.cn</email>
</ distributor >
```

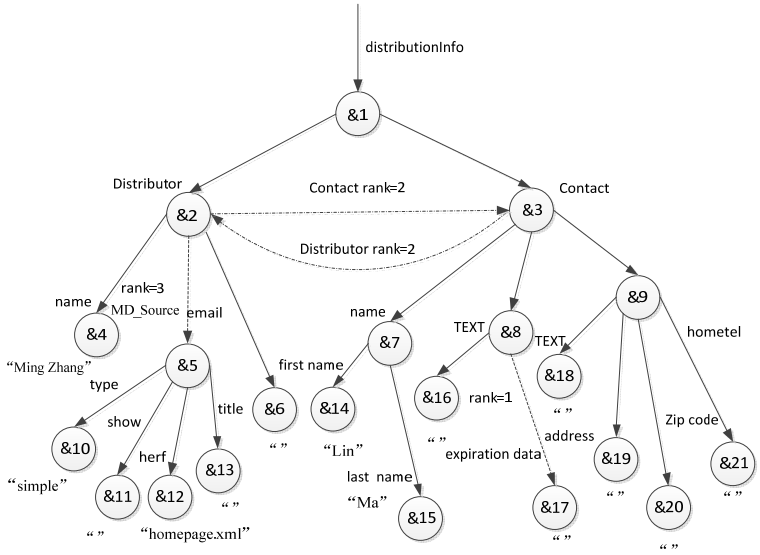

Figure 4. X-3DCDM model's directed graph

```
< contact='3' distributor id='2' >
<!- Internal IDREF reference, and rank has a value of 2- >
<name>
<firstname>Lin</firstname>
<lastname>Ma</lastname>
</name>
<IDcard expirationdate='2015/05/03' >25003408</IDIcard>
<memo>
<!-Hybrid content, and a node is individually created for the character
data, where the edge is labeled as TEXT->
here is some information of this contact
<address>No.3 Shang Yuan Cun,Hai Dian District Beijing</address>
<zipcode>100044</zipcode>
<hometel>010-5146****</hometel>
</memo>
</contract>
</distributioninfo>
```

Adopting the X-3DCDM as the public data model in Railway Distributed System Integration is not only capable of flexible expression of semi-structural data such as XML and HTML documents etc., but also is convenient to realize transformation from and to traditional data models, including relationship models and object-oriented models.

VII. CONCLUSION

The goal of sharing railway basic data is to integrate discrete data resources in order to establish service-oriented data sharing services architecture. This architecture provides optimum designing and operational management of high-speed railway transportation by distributing and centralizing data. To meet the integrated access requirements for railway data, the architecture should include the features of location transparency, name transparency and pattern transparency in heterogeneous distribution circumstances. In this paper a common metadata dictionary pattern is designed to meet the multiple requirements. It is composed by three main parts--- representation of data source network location, data schema and data content. In order to provide support for dealing with heterogeneousness of the data models among data sources in railway distributed system integration, an analysis and comparison of popularly used data models based on the graph is given in the paper on the basis of railway data characteristics analysis. And an XML-oriented 3D common data model (X-3DCDM) is brought forth after comprehensively considering and optimizing the advantages of the existing models. Based on the data model with a root connected directed graph, the model fully takes into account the layers relationship and flexibility of syntax expression in data integration, realizing its mapping to other data models including the relationship model and object-oriented model, so that it functions as the role of some type of metadata, which can describe the data integration consistently, unify the algebra operations of data integration and effectively support the data integration of railway distributed systems. A higher level of integration for railway distributed systems such as the process integration needs a still higher level of conception and abstraction. For example, a deep study into the Ontology, which features the establishment of a set of shared terminologies and information expression structure. With the help of Ontology, that heterogeneous information between multiple data sources can become homogeneous information, thus effective process integration can be achieved.

ACKNOWLEDGMENT

The authors gratefully acknowledge the editor and anonymous reviewers for their valuable comments and constructive suggestions. This research was supported by the Beijing Municipal Science & Technology Commission key project (Z090506006309011), National Science and Technology Support Program topics (2009BAG12A10), the State Key Laboratory of Rail Traffic Control and Safety (RCS2009ZT007), Beijing

JiaoTong University, and partially supported by the MOE key Laboratory for Transportation Complex Systems Theory and Technology School of Traffic and Transportation, Beijing JiaoTong University. .

REFERENCES

[1] Xiaoyun CHEN. Chinese Railways, vol.2, 2002, pp.36-40.
[2] Jiangang MA, Tao HUANG, Jinling WANG, Gang XU and Dan YE. "Underlying Techniques for Large-Scale Distributed Computing Oriented Publish/Subscribe System," Journal of Software, vol.17, 2006, pp.134-147.
[3] Gang XU, Tao HUANG, Shaohua LIU, Dan YE. "Survey on the Core Techniques of Distributed Application Integration". Chinese Journal of Computers, Vol.28, 2005, pp.433-442.
[4] Xin Dong, Alon Y. Halevy, Cong Yu., "Data Integration with Uncertainty," in proceeding of VLDB September, 2007, Vienna, Austria, VLDB Endowment, ACM Press.
[5] Nash A., Huerlimann D. "RailML – a standard data interface for railroad applications", Computers in Railways IX, WIT Press, Southampton, 2004. pp. 233-240.
[6] S.Suzuki and H.Ando. "A modular network scheme for unsupervised 3D object recognition", Neuro computing, vol.31,2000, pp15-28.
[7] Smith, M.J. "Digital elevation models for research: UK data sets, copyright and derived products." Elevation Models for Geoscience. London, Special Publications, vol.345, 2010, pp129-133.
[8] Carlo DELL'AQUILA, Ezio LEFONS. "Design and Implementation of a National Data Warehouse" Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006, pp342-347.
[9] Brian L. Smith. "Investigation of Extraction, Transformation, and Loading Techniques for Traffic Data Warehouses". Transportation Research Record: Journal of the Transportation Research Board. vol.1879. 2004.
[10] Jin Ma."Managing metadata for digital projects Library Collections,"Acquisitions,& Technical Services,vol.30, 2006, pp.3-17.
[11] Ning FU, Xingshe ZHOU, Tao ZHAN. "Data models for information integration ",Application Research of Computer,vol.25(5),2008(5), pp.1285-1294.
[12] Jianghong HAN, Shuli ZHENG. "Research on Web Data Model Oriented to XML", MINI-MICRO SYSTEMS, 2005(4), vol.26 (4), pp.609-613.
[13] Zhenying He, Jianzhong Li, Chaokun Wang. "A Data Model for XML Database," Journal of Software, 2006, pp.759-769.

Hanning Wang received her M.E degree from the PLA Information Engineering University, China, in 2006. She is now a Ph. D. Candidate of State Key Laboratory of Rail Traffic Control and Safety, Beijing JiaoTong University, China. Her current research interests are in area of data mining and distributed computing.

Weixiang Xu received his PhD from the Beijing JiaoTong University. He has a wide range of research experience and interest in information systems and data mining. He is now a Professor, tutor of a Ph.D. student at the Beijing JiaoTong University, China.

Chaolong Jia received his B.E degree from Liaoning Technical University, China, in 2003. He is now a Master-Doctor combined program graduate student of State Key Laboratory of Rail Traffic Control and Safety, Beijing JiaoTong University, China. His research interests are deformation and structal safety, fault prevention technology, technical support, information systems and data mining in railway.

About corresponding author:
Name: Hanning Wang
Address: State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, No.3 Shang Yuan Cun, Hai Dian District Beijing, China Post-Code 100044
Tel: 0086+18600021363
Email: 06114158@bjtu.edu.cn