XML Schema Design and Management Guide

Part I: Overview

Version 0.3.01

13 June 2003

Produced by the Center for E-Commerce Infrastructure Development
The University of Hong Kong

Status of This Document

This document is in the draft status and is pending review by the XML Coordination Group (XMLCG). The XMLCG has been convened by the Information Technology Services Department of the Hong Kong SAR Government (HKSARG) to develop strategies to facilitate more effective adoption of XML in the Government. Suggestions to be received from the XMLCG will be considered and will be used to revise this document accordingly. The XMLCG-reviewed version will be distributed to all Government bureaux and departments (B/Ds) and their IT contractors for consultation. Suggestions from the consultation will be considered and used to develop the final version.

The final version will be published on the Internet for sharing among B/Ds and their IT contractors.

This version is: **XML Schema Design and Management Guide - Part I: Overview**, Version 0.3.1 (13 June 2003).

Table of Contents

1. Introduction

1.1. Objectives of this Guide

A key strategy of our e-government initiative is to develop joined-up e-government services. The necessary conditions for implementing joined-up services are:

- The business partners' commitment to integrate their business functions electronically with a view to deriving more business value through the collaboration;
- **Process interoperability**: agreement on how the business activities of the concerned parties affect each other, i.e. the business rules; e.g. when the seller receives a purchase order from the buyer, the seller should accept or reject this order within a specified period;
- **Data interoperability**: agreement on what information has to be transmitted from one party to another, and the definition and representation of such information; e.g the "delivery date" has to be specified on a purchase order, and "delivery date" means the date on which goods shall be received by the buyer, and the representation of the "delivery date" adopts the ISO 8601 standard[1]; and
- **Technical interoperability**: agreement on what communication protocol and message format to be used when one party sends information to another; e.g. the purchase order shall be encoded in XML, as defined by XML Schema, and XML Encryption and XML Signature shall be applied on certain content components, and the XML message shall be sent via HTTP.

The HKSARG Interoperability Framework (IF) is an enabler for the implementation of joined-up e-government services. It facilitates technical interoperability and data interoperability among B/Ds and their business partners. This is achieved by promulgating technical standards, and by putting in place a mechanism for the central alignment and specification of the **data elements** that has potential for reuse in multiple joined-up projects. A data element is a piece of information for exchange and the concepts of data elements are discussed in Section 6.

The technical standards for facilitating technical interoperability are specified in the IF document (http://www.itsd.gov.hk/itsd/english/infra/eif.htm). The standards are being reviewed every 6 to 12 months, in order to keep in pace with technology advancement and to take into account new interoperability requirements.

This guide facilitates data interoperability by providing:

- A methodology for business analysts to specify the definitions and representations of information in a consistent and structured way as reusable information models;
- An approach for programmers to convert the information models (or specifications) of the data elements into XML Schema Definition (XSD) code pursuant to the IF's

[1] ISO 8601 is the standard on representation of dates and times published by International Organization for Standardization.

recommendation of using XML for information exchange and XML schema for defining the structure, content and semantics of XML documents;

- Guidelines for the central alignment of the data elements that has potential for reuse in various joined-up services, thereby standardizing the XSD code for these data elements; and
- Guidelines for project teams to adopt suitable centrally aligned data elements and their standardized XSD code and also contribute reusable data elements for central alignment.

1.2. Structure and Audience of this Guide

This Guide comprises 4 parts in total.

PART I: Overview

This **Overview** is primarily for e-government project teams, project owners and all the parties involved in the central alignment of data elements to get a high-level understanding of the overall mechanism.

PART II: XML Schema Design Guide

The **Design Guide** is intended for business analysts responsible for modelling the business process and information requirements in a project. It also provides guidelines for programmers to convert the information models specified by the business analysts into XSD code.

PART III: XML Schema Management Guide

The **Management Guide** serves as a handbook for the parties involved in the central alignment of data elements for developing reusable XML Schemas. It also enables project teams to have a better understanding of how data elements are centrally aligned, in particular the role of project teams as a contributor of reusable data elements for central alignment.

PART IV: Appendices

The appendices provide supplementary information to help the readers understand this Guide. They include:

Appendix 1 – Case Study – Application for Import and Export Licences for Pharmaceutical Products and Medicines

Appendix 2 – Implementing eBusiness Solutions

Appendix 3 – Intellectual Property Rights of Registry Artifacts

Appendix 4 – Recommended List of CCTS

Appendix 5 – Core Component Type Worksheet

Appendix 6 – Sample XML Schema Design Worksheets

Appendix 7 - Glossary

2. The Data Interoperability Problems Faced by Project Teams Today

Some of the data interoperability problems faced by project teams include:

- **Incompatible data definitions**: e.g. the shipment delivery date could be interpreted as the date on which a customer obtains the product or the date on which a product is unloaded at the intended destination, which could be different; and
- **Incompatible data representations**: e.g. the postal address is the information that needs to be exchanged; the sender could be maintaining postal address as 5 lines of 35 characters each whereas the receiver could be maintaining postal address as 4 lines of 50 characters each.

The parties planning to exchange information must conduct a data alignment exercise before they can accurately exchange information. They need to agree on the definition and representation (i.e. structure, permissible values, etc.) of each data element to be exchanged.

And when they implement the information exchange interface, either or both parties may need to perform data conversion or data mapping between the exchanged data and the data maintained in their internal systems.

Traditionally, to implement a joined-up service that involves multiple parties, the concerned project teams perform data alignment with the sole purpose of satisfying the project requirements from these parties. In case a new party needs to be involved in the joined-up service some time after the service has been in production, the data alignment exercise may need to be repeated and the data conversion software may need to be rewritten to satisfy possibly different requirements from the new party.

Such repetitive data alignment and data conversion efforts should be minimized as far as possible.

3. Data Interoperability Strategy

The problems mentioned above, namely repetitive data alignments and data conversion, can be minimized by the central alignment of data elements (and the standardization of XSDs for these data elements), particularly for those data elements that are often involved in data exchanges between B/Ds.

E-business consortiums around the world have initiated similar XML standardization efforts. Examples include:

- The development of e-business standards by the OASIS Universal Business Language (UBL) Committee;
- The development of supply chain management standards by the EAN International and the Uniform Code Council, Inc.;
- The development of business reporting standards by the eXtensible Business Reporting Language (XBRL) Consortium, etc.

While having two or three parties agree on the definition and representation of one data element may not be an easy task, having over 90 B/Ds agree upon the definitions and representations of a collection of data elements is even harder, given that each B/D has its own legacy systems using different data formats.

B/Ds must recognize the need for interoperability standards and commit themselves to the adoption of existing standards and the development of additional standards based on business needs.

In respect of the central alignment of data elements, B/Ds may involve themselves by joining the XML Coordination Group (XMLCG). The XMLCG is the top-level consensus making body responsible for the central alignment of data elements. B/Ds may also nominate Common Schema Liaison Officers to provide requirements and comments in relation to the central alignment of data elements.

And when project teams develop project-specific XML Schemas, they should observe the following data interoperability measures:

1. Adopt industry standards where appropriate;
2. Design quality and reusable Project Schemas by applying the schema design methodology provided in this Guide (Project Schemas are information models and XML Schemas developed by the project team for project-defined data elements. The concepts of Project Schemas are discussed in detail in Section 6);
3. Adopt Common Schemas in projects whenever possible (Common Schemas are information models and XML Schemas developed for the centrally-aligned data elements. The concepts of Common Schemas are discussed in detail in Section 6);
4. Share Project Schemas with other project teams; and
5. Contribute reusable data elements from Project Schemas for creation of new Common Schemas

An overview of the above measures is provided in the next section.

4. Data Interoperability Measures

4.1. Adopt industry standards where appropriate

Before project teams design Project Schemas for implementing a joined-up service, they should first consider adopting industry standards where appropriate. For example, if they are defining an interface for procurement, they should study existing and emerging industry standards such as UBL to see if the standard can fulfill their project requirements or whether their business practices can be reengineered to align with the standard practice promoted in the industry standard. One resourceful Website of published industry standards is http://www.xml.org. The business analysts with domain-specific knowledge should be well aware of the development of industry standards for their specialized industry sectors. Only when suitable industry standards do not exist should they attempt to design their own Project Schemas.

4.2. Design quality and reusable Project Schemas

When no suitable industry standard is available, project teams are recommended to apply the design methodologies specified in Part II of this Guide for designing Project Schemas when they implement joined-up services. The project teams shall apply the methodologies during the system analysis and design (SA&D) stage of the system development life cycle.

Part II covers a **business process modelling (BPM)** methodology, which project teams may **optionally** adopt to model the collaborative business functions and to identify the business documents involved in the collaboration. Part II also covers a **business information modelling (BIM)** methodology, which project teams are **highly recommended** to adopt to model the business documents involved in the collaboration.

Specifically, the BIM methodology facilitates business analysts in the project team to decompose a business document into modular data components, which serve as the building blocks for reassembling various business documents. Complex data components are decomposed into simpler components. For example, a purchase order is a complex data component, which can be decomposed into simpler components such as buyer contact information which appears on other business documents like delivery note, invoice, etc. These data components are referred as data elements in this Guide. Finally, a hierarchical structure of data elements is formed to represent the business document. These data elements are specified as **information models,** covering the definition, representation, etc. that accurately reflect the data element's attributes. Such decomposition of complex data elements into modular (and simpler) data elements enhances the reusability of the data elements. At the same time, the careful specification of data elements (as information models) avoids ambiguity and enhances the shareability of these data elements across different parties.

The specifications of data elements need to be converted into XSD code. Part II of this Guide also provides an approach for programmers to translate the information models into XSDs.

The information model and the XSD of a project-defined data element together form the Project Schema for that data element.

4.3. Adopt Common Schemas whenever possible

When project teams develop Project Schemas, they are required to adopt the Common Schemas that are considered mature whenever possible. Common Schemas are published in the **Central Registry** to facilitate adoption by project teams to develop Project Schemas. (The concept of the Central Registry is discussed in Section 6.)

While the business analysts are performing business information modelling, they should search the Central Registry for suitable Common Schemas. The suitability of a Common Schema is determined by whether the data element specification of that Common Schema (specifically the definition, representation, and business contexts) meets the requirements of a particular data element identified in a business document.

Besides, each Common Schema is associated with a maturity level. The business analysts should adopt any suitable Common Schemas that are considered to be mature.

When suitable Common Schemas are located, the data element specifications of these Common Schemas, together with the XSDs, shall be imported to the **Project Registry** for developing Project Schemas. (The concept of the Project Registry is discussed in Section 6.) As a result, the Project Schemas of the business documents will incorporate the suitable Common Schemas.

There may be cases where customization is needed before a Common Schema can be used in a project. In such cases, the business analyst may perform customization as necessary and at the same time try to maintain the reusability of the customized schema as far as possible.

The reuse of the Common Schemas significantly minimizes repetitive data alignment efforts and enhances the data interoperability for the Project Schemas designed for use in joined-up services.

When a business analyst decides to reuse a Common Schema, he should register in the Central Registry that his project is reusing that Common Schema. The registration of schema reuse allows the project team to be notified of any subsequent change to that schema. The reuse statistics also enable the assessment of the maturity level of the Common Schemas.

When project teams define their Project Schemas, apart from adopting relevant Common Schemas, the business analyst should also try to identify suitable data elements / schemas in industry standards and other e-government projects for adoption.

4.4. Share Project Schemas with other project teams

Project teams should organize information models and XSDs of project-defined data elements (i.e. Project Schemas) in a Project Registry for subsequent reference during the system development and maintenance life cycle. Project Schemas are an integral part of a project's system documentation, just like design specification, source code, etc.

Since Project Schemas may affect a system's future integration with the systems of other B/Ds and external parties, project teams are recommended to share Project Schemas with other B/Ds and external parties where relevant.

Such sharing also allows other project teams working on similar initiatives to share best practices and reusable schemas, thus maximizing the reuse of schemas.

To facilitate the sharing of Project Schemas among e-government project teams, project teams are recommended to register their projects on a centrally maintained list of projects with their Project Schemas openly accessible. This list provides links to the Project Schemas and other information of various joined-up service projects for reference by all parties.

4.5. Contribute reusable data elements for creation of new Common Schemas

Although the sharing of Project Schemas among project teams can, to a certain extent, facilitate the reuse of schemas, it is difficult for a project team to locate useful Project Schemas from various other projects' Project Registries, given that each Project Registry may be organized in a different manner. A more effective way to facilitate schema reuse is to have them centrally aligned as Common Schemas. These Common Schemas can cover both data elements that are applicable in all business contexts, and domain specific data elements that are only applicable in specific business contexts.

The success of a data interoperability strategy is measured by the adoption rate of Common Schemas, and the volume of Common Schemas. The more Common Schemas there are, the more we are in a position to ease project teams' efforts in data alignment and modelling, and schema design.

The most effective way to collect reusable data elements to create new Common Schemas is through contribution by project teams. Project teams are highly recommended to analyze their Project Schemas and contribute the project-defined data elements that have potential for reuse by other projects for central alignment. These may be data elements newly defined by the project, or they may be data elements customized based on an existing Common Schema.

The data elements contributed by project teams will be aligned centrally in accordance with the management mechanism proposed in Part III of this Guide. The process of aligning a data element centrally will take some time, given that the alignment process may involve over 90 B/Ds in reaching consensus on the data requirements. When the central alignment process completes, the information model shall be specified for the data element and the XSD code shall be developed based on the model.

The information model and the XSD of a centrally aligned data element together form the Common Schema of that data element. The contributing project team is highly recommended to adopt the new Common Schema in its project.

It may be too ambitious trying to accommodate all schemas used in government services as Common Schemas, given the difficulty in doing data alignment across a user group as large as over 90 B/Ds. We can start with data elements that are easier to get aligned, and then build up experience and confidence in the process.

4.6. Highlights of measures from a project team's point of view

The flowchart shown in Figure 1 highlights the steps that project teams should follow when they develop XML Schemas in a joined-up service project, in the perspective of facilitating data interoperability.

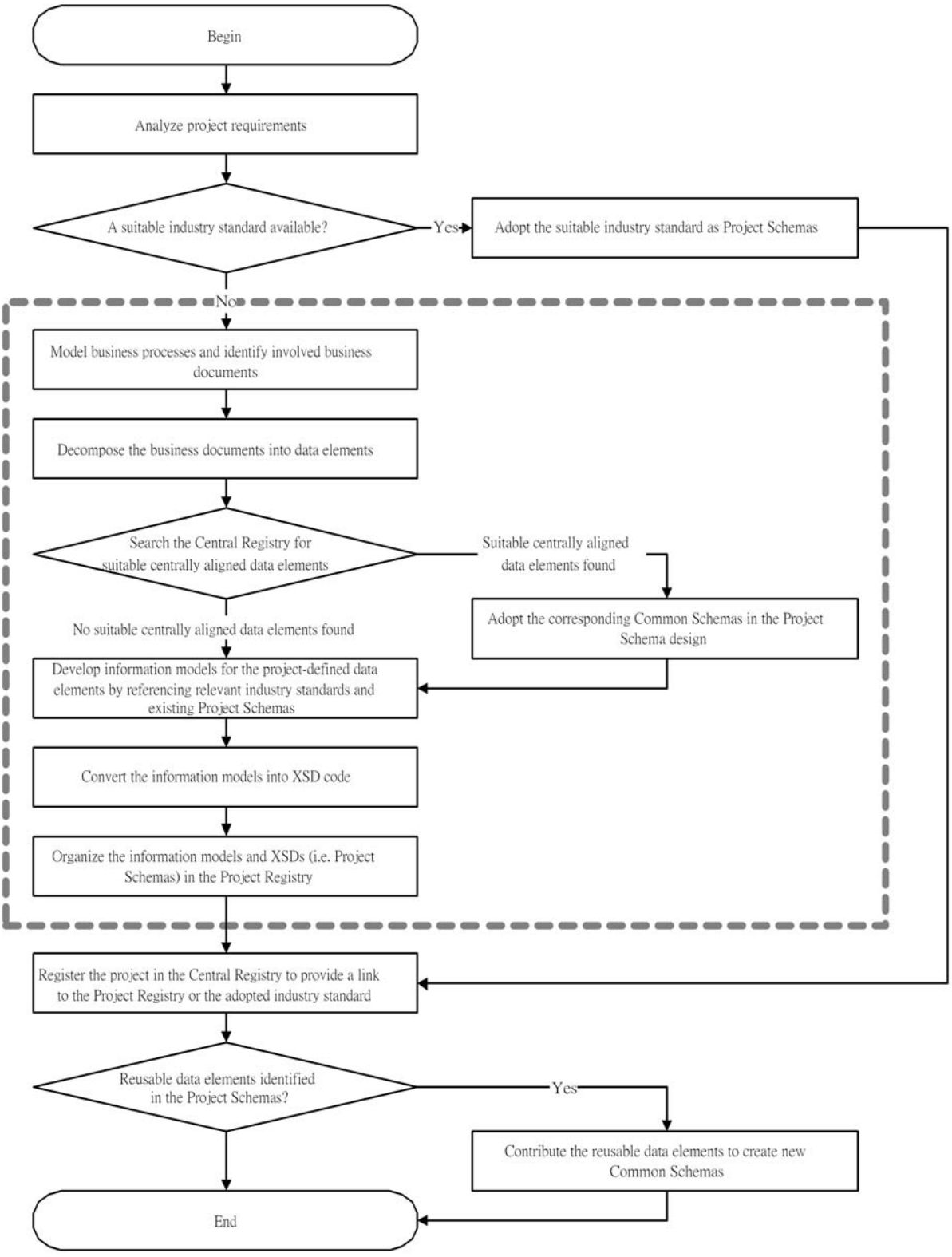

Figure 1: Steps in developing XML Schemas in a joined-up service project.

5. Guiding Principles Governing the Data Interoperability Measures

The guiding principles behind the measures described in the previous section are as follow:

- The mechanism should facilitate the progressive central alignment of data elements and standardization of the corresponding XML Schemas for reuse across all B/Ds;
- The mechanism should facilitate the location and retrieval of reusable Common Schemas for adoption by project teams in developing Project Schemas;
- The mechanism should minimize inconveniences, overheads and delays that may be induced on projects;
- The mechanism should encourage project teams to contribute reusable data elements for central alignment and the alignment process should be completed within a reasonable time frame so that the contributing party can adopt the aligned data element;
- The mechanism should encourage B/Ds to involve themselves in the creation of Common Schemas;
- The mechanism should facilitate B/Ds to build consensus on the definition, representation, and usage of data elements when creating Common Schemas;
- The mechanism should not induce excessive overheads on B/Ds for the central alignment of data elements (e.g. efforts required from their Common Schema Liaison Officers and representatives in XMLCG);
- The mechanism should minimize additional training required to enable the various stakeholders (i.e. business analysts, programmers, Members of the XMLCG, and the Common Schema Liaison Officers) to support the mechanism;
- The mechanism should promote international best practices in the alignment and the specification of data elements (and the development of reusable XML Schemas) with a view to enhancing the shareability of Common Schemas across systems and environments; and
- The mechanism should be transparent; all decisions should be well documented and openly accessible, and the stakeholders should be kept informed of the execution status.

6. Key Concepts

This section recaps the key concepts mentioned in this Guide. The Glossary in Appendix 7 also provides definitions of terminologies used in this Guide.

6.1. Joined-Up Service

A joined-up service is the interconnection of business functions operated by different business partners electronically through the exchange of business information to perform a collaborative business process.

Usually, the business partners first reengineer the entire workflow covering all parties with a view to streamlining the operation. The business partners also have to agree among themselves the collaboration model, which includes:

- The process interaction model (how the business function of one party interacts with and affects the business function of other parties);
- The information exchange interface (what information needs to be exchanged in each interaction and the meaning and representation of such information);
- The technical specifications to be followed for sending information from one party to another; and
- Other details like response time, service level agreements, etc.

For some specific joined-up services where a generic business process pattern can be identified, the industry has defined standards for process interaction, information exchange, and technical interaction. Under these situations, B/Ds are recommended to adopt the industry standard if it can meet the project requirements.

Once the collaboration model has been agreed upon, the business partners then automate their own business functions based on the agreement. Finally all automated processes/services are interconnected to collaborate electronically through the exchange of business information.

In the context of this Guide, the stakeholders of a joined-up service include the following:

Business partner: a B/D or an external organization that provides and/or receives business information in a joined-up service. Each business partner is responsible for offering the automated services that implement the business functions it is responsible for, in accordance with the collaboration agreement made among all the business partners.

Project team: the team involving business users, domain experts, business analysts, programmers, and so on that designs and develops the automated services that implement the business functions responsible by a business partner of the joined-up service.

6.2. Project-defined data elements versus centrally aligned data elements

A business document involved in an information exchange process usually comprises many data components, some of which may be further decomposed into simpler data components. For example, a purchase order sent from the buyer to the seller is comprised of buyer contact information, delivery information, goods ordered and quantity, and so on. The buyer contact

information may be further decomposed into buyer's name, correspondence address, telephone number, and so on.

These data components, regardless of whether they are simple data components like a telephone number, or complex data components like the entire purchase order, are referred as **data elements** in the context of this Guide.

Some data elements are repeatedly used in many business documents. For example, the buyer contact information may appear on the purchase order, the delivery note, the invoice, etc.

In order to ensure that all business partners have the same interpretation on a data element, the data element should be carefully specified with information such as its definition, identification, representation (such as the structure, format, permissible values, etc.), usage rules, etc.

In the context of this Guide, the data elements used in a particular joined-up service are referred as **project-defined data elements**. In cases where the business partners in a joined-up service have decided to adopt industry standards at their collaboration interface, then the project-defined data elements will correspond to the data elements defined by the industry standard.

Some data elements are used for exchange in various joined-up services. For example, the "Hong Kong Identity Number" may be used in many joined-up services for identifying a person. In order to avoid repetitive data alignment and data conversion, and to save different project teams' effort in re-designing the representation of the same data element, a mechanism will be put in place for the central alignment of data elements.

The process for the central alignment is described in Part III of this Guide. Those data elements that have been agreed upon through the central alignment process are called **centrally aligned data elements**.

Besides the definition, identification, representation and usage rules mentioned above, each centrally aligned data element is also associated with business contexts. A business context defines and confines the business situations in which the data element should be applied.

The definition, identification, representation, and business contexts of a data element shall be specified according to the BIM methodology provided in Part II of this Guide as the metadata of that data element.

The central alignment of data elements will be an on-going and strategic exercise. Initially, those data elements that are often involved in data exchanges between B/Ds will be aligned first. In the long run, the major source of data elements for central alignment should come from the joined-up services themselves. Project teams are highly recommended to select project-defined data elements that has potential for reuse by other joined-up services and submit these data elements for central alignment.

6.2.1. Conflicts between industry standard data elements or between an industry standard data element and a centrally aligned data element

It is possible that the various industry standards may overlap on some data elements, or a data element in an industry standard may overlap with a centrally aligned data element. If two business parties have previously adopted their corresponding industry standards and subsequently identified incompatible data attributes when they try to interconnect their business functions electronically, then data conversion or data mapping will have to be performed by one party. Such inconvenience is likely to remain in the foreseeable future.

Although such inconveniences may not be avoided in some situations, the trend to develop industry standard is a right direction to go. The more convergence B/Ds can achieve, the

more efficient B/Ds will become in e-business integration, and the more it is likely to have "standard converters" (or adapters) for translating from one industry standard to a related industry standard.

6.3. Project Schema versus Common Schema

In principle, the carefully specified data elements are not bound to any particular syntax. They can be bound with specific message syntax such as XML or UN/EDIFACT, etc. In accordance with the IF, XML should be used as the default format for information exchange between B/Ds or between a B/D and an external party; and XML schema should be used as the default schema language to define the structure, content, and semantics of an XML document.

Part II of this Guide provides guidelines for converting the information models of data elements into XSD code.

The XSD that corresponds to a project-defined data element together with its information models are referred as **Project Schemas**.

The XSD that corresponds to a centrally aligned data element together with its information models are referred as **Common Schemas**.

Each Common Schema is associated with a maturity level. This maturity level indicates how stable a Common Schema is with respect to change. The more mature a Common Schema is, the less likely that Common Schema will be changed. The rules for determining the maturity level is described in Part III of this Guide.

When project teams implement the information exchange interface between B/Ds or between a B/D and an external party, they are **required** to adopt the Common Schemas that are considered mature for adoption if the Common Schema matches the project requirements, taking into consideration the data definition, representation, business contexts, usage rules, etc.

6.4. Project Registry versus Central Registry

The registry provides a way for organizing information. In the context of this Guide, the information that needs to be organized includes:

- Information models, XSDs and other administrative information for data elements
- Controlled vocabularies or code lists that provide sets of permissible values for the data contents (e.g. the department code of B/Ds) or metadata (e.g. the permissible values used to describe the business contexts of a data element).

The registry, in the context of this Guide, provides an organized way for project teams (in particular business analysts) to locate suitable data elements for reuse. Since the registry is mainly for human use, it should provide a convenient user interface. The sophistication of the user interface very much depends on the volume and nature of information being maintained in the registry. Programmatic interfaces such as UDDI or ebXML Registry Service are considered beyond the scope of this context.

In the registry, the information models and XSDs of data elements can be organized in the form of a **Data Dictionary**. In other words, the registry contains, at least, a Data Dictionary and a collection of code lists.

Project teams should organize the Data Dictionary of project-defined data elements (i.e. Project Schemas) in a **Project Registry** for subsequent reference during the system

development and maintenance life cycle. Project Schemas are an integral part of a project's system documentation, just like design specification, source code, etc.

Since Project Schemas may affect a system's future integration with the systems of other B/Ds and external parties, project teams are recommended to share Project Schemas with other B/Ds and external parties where relevant.

Such sharing also allows other project teams working on similar initiatives to share best practices and reusable schemas, thus maximizing the reuse of schemas.

To facilitate the sharing of Project Schemas among e-government project teams, project teams are recommended to register their projects on a centrally maintained list of projects with their Project Schemas openly accessible. This list provides links to the Project Schemas and other information of various joined-up service projects for reference by all parties.

With regard to the data elements that have been centrally aligned and carefully specified, they shall be maintained in a **Central Registry** for access by all B/Ds and their business partners. The list of projects with openly accessible Project Schemas mentioned above will also be maintained in the Central Registry.