Three Myths about Canonical Data Models

In modern IT world, many people are busy with systems integration, and in the context of that, with *canonical data models*[1] (CDM's). The requirements that are set on companies nowadays, by clients, regulatory organizations and business partners like suppliers, demand effective communication processes, both internal and external to the organization, and a flexible IT.

A CDM can help meet these requirements. A condition for succes, however, is for the concept to be used properly. This article argues that the concept of a CDM is often misunderstood, as a result of which the odds are high for an ineffective application of it. The article hopes to contribute to a discussion about a correct setup and use.

Myth 1: The CDM is merely a means to loose coupling

In the literature, attention is often restricted to one specific application of CDM's: the use of the CDM as a means for *loose coupling*[2] in a middleware product like an *Enterprise Service Bus (ESB)*. This is without doubt an interesting application, that can greatly improve IT flexibility in an organization. However, the potential of a CDM is much higher than only this use. In a random order, a CDM can also be used as:

Data Catalogue. In order to support the reuse of business data, the CDM gives insight into which data are available, where they are, how they're structured, and possibly even into their quality and into the rules that apply about their authorization, classification and privacy.

Semantic Data Model and *Common Business Language.* An important cause of miscommunication is what is called the *semantic problem*; miscommunication by parties not being aware of the fact that they are interpreting messages in different manners. The CDM offers standardized interpretations (definitions)[3] of terms (CDM as a Common Business Language) and of data representations (CDM as a Semantic Data Model). The CDM can shed light on existing semantic differences (semantic disharmony) that can lead to communication problems, and it can help solve these problems.

Data Model Catalogue. To support the reuse of data models, the CDM, as a backbone model in a modeling maintanance environment, offers a set of data models of a certain domain and a certain status. These models can have multiple functions, for instance in the design of new functional components.

---

[1] A Canonical Data Model is a data model that is meant to have or acquire a business-broad coverage, and to be as commonly accepted as possible (hence 'canonical'), within the confinements of the ambitions of choice and available means.

[2] Loose coupling is the coupling of functional components with minimal interdependencies. By tuning the components on the WHAT is being communicated only, for instance (that is, on the data semantics), one deminishes the amount of interdependency between components. The consequence of this is a great flexibilization of IT.

[3] An important activity in realizing systems communication is creating *data mappings*. Data mapping is a process that not only is a matter of technically knitting attribute types togehter and defining technical data transformations and –translations between attribute types. An underestimated aspect of this mapping process is to determine which attribute types can be knitted together in a *semantically meaningful* way in the first place, and which *semantical transformations* are necessary in doing so. Here, the semantical definitions of the CDM come in handy.

Thermometer In the IT. To be helpful in improvements trajectories, the CDM gives an insight into where in IT bottle necks and opportunities exist. One might think of unbridgeable data disharmony, for example, or of unwanted double data registrations, gaps in data sets or insufficient data quality.

Dedicated models can be created for these applications, but it is not adviseable to make and maintain multiple, overlapping, company-wide data models. That's a very costly situation that will lead to many inconsistencies as well. The CDM pretents to be that one data model that can support all of these applications.

Myth 2. External Standards in the CDM

An often heard idea about CDM's is, that in creating a CDM, it would be a best practice to implement external data or communication standards one-to-one and in full extend in your CDM.

CDM's and external data or communication standards both are sets of agreements, 'standards', that can be put to work in a 'loosely coupled' communication process. However, there are differences as well. Table 1 gives an oversight.

Table 1. Differences between Canonical Data Models (CDM) and External Standards (ES)

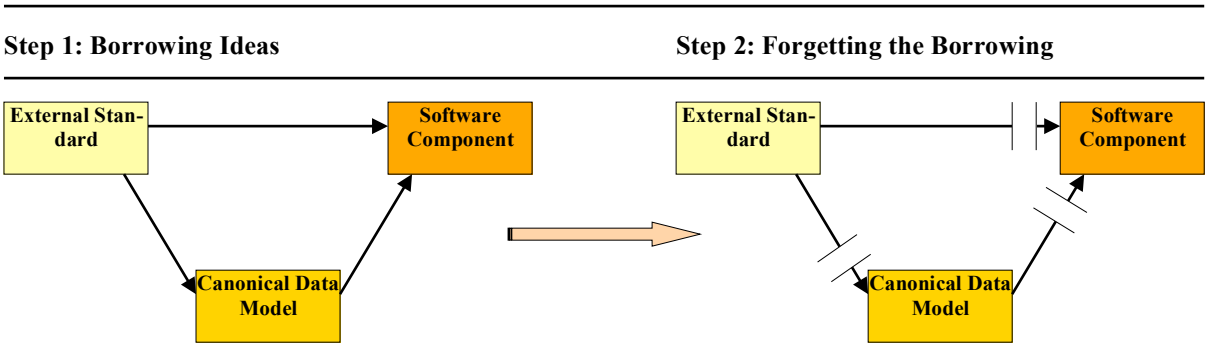
aspect	CDM	ES
objective	advocacy of the organization itself (including competitive position)	advocacy of communication between organizations (including cooperation)
primarily directed to	communication within the organization	communication between organizations
desired decoupling	between components within the organization	between components of organizations in a partnership
standardizing organization	the organization itself	external standardization institute
ownership	the organization itself	external standardization institute
influence of organization	big	limited
reach	the organization	space between organizations (or parts)
scope of the *loose coupling* mechanism	(in the middleware) between systems of a single organization	(in the middleware) between systems of multiple organizations

These differences require for the respective standards to be applied differently. The purpose of external standards is to improve the inter organizational communication, not to internally standardize your organization. This idea is similar to the objective of the artificial language Esperanto, which was not to replace existing languages, but to provide people with a second language, a shared standard language for communication between people who speak different mother tongues.

Because external standards are not aimed at looking after your own organization's specific interests, they should not be implemented 1:1 in a CDM just like that. The only correct way to use external standards for constructing a CDM is to identify interesting parts in it, screening it for desirability and consistency with other applicable regulations, and to build it in the CDM as a 'borrowed idea'. This should be done in accordance with current construction principals

that apply to the CDM. Done this way, the organization can preserve its own self-determination and identity, and it can determine and realize its own competitive strategy[4].

After borrowing an idea, all dependency relations between the external standard and its implementation in the CDM should be cut immediately (see Picture 1). Preserving these dependencies will stand in the way of self-determination and flexibility. Changes made to the external standard or a choice for another standard must be implemented in the CDM as well. Moreover, the CDM can only be changed when the external standards change accordingly, which is something one cannot enforce, normally.



Picture 1: The principle of borrowing and forgetting

Complications in the use of external standards as a CDM- building block

There are two more reasons why external standards should not and cannot be implemented directly in a CDM.

Mutual inconsistency. Standards can be inconsistent with each other. Therefor, one can never blindly build them into a CDM. One cause is, that many standards actually standardize on more than one dimension simultaneously, e.g. besides the actual data, also the structure of the messages they are contained in, the language chosen (e.g. XML), and the way modeling is done in it ('best practices XML'). One possible solution is to 'borrow' single ideas from an external standard that do not cross a dimension border.

Weak semantic standardization. External standards often are not very strong at standardizing the intended interpretation of data elements (semantics). Hence, they are often multi-interpretable themselves. This shows in the fact that their users interpret and use them in their own way, in a semantic sense, despite the fact that they are 'standards'.

Myth 3. Standardizing Components with the CDM

This myth has it, that is would be wise to use the CDM to standardize functional components (applications, web services, databases, and so on).

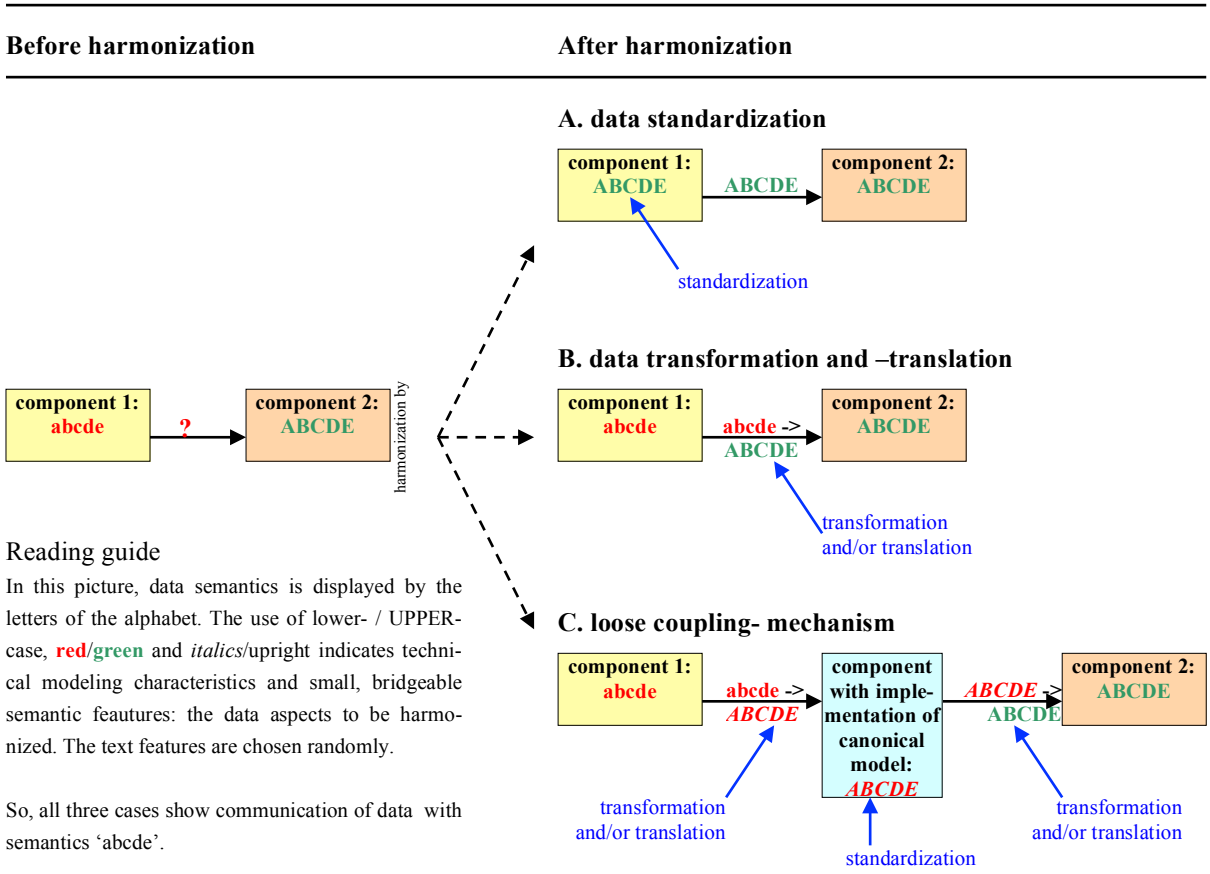
A CDM, or part of it, can be used as a data model for the design of an automated IT component. Thereby, one can have a design for a component to be purchased, or to be build, available rather quickly and cheaply. In principle, reuse of data models (see myth 1) is a good thing.

[4] It is wise though to implement external standards 1:1 and to a full extend for the purpose of *external* communication. After all, it was agreed with external business parties to communicate in the standardized manner. This is how external standards are meant to be used.

But, the relationship between the CDM and the implementation of the borrowed ideas in the components should be cut immediately after, in order to decouple both. Otherwise, the dependancy will lead to IT inflexibility, something you wanted to get rid of in the first place.

Another goal one can have in mind for the reuse of a CDM is *data harmonization*. Components can only communicate well, to the extend that they are in harmony with each other. This means that between the components there are either no differences in data semantics, structure and format, or such differences can be bridged during the communication proces.

These two possible situations can be reached applying two different harmonization strategies: *data standardization* and *data transformation and –translation* (see Picture 2, parts *A* and *B*).



Picture 2: Three different strategies for data harmonization

The essence of the loose coupling concept is, to embrace the idea of 'local freedom' and central functional tuning with full conviction. In fact, you want local parties to have the freedom to look after their own interests, without them being hindered too much by communication parties. Tuning with communication parners therefor has to be restricted to functional matters, i.e. data semantics and communication patterns. The goal of this concept is to gain flexibility.

If you apply the CDM as a means to standardize the decentralized components, you simply ignore this concept. You are not embracing the idea of loosely coupling components. The result will be an inflexible IT.

A benefit of internal component standardization can be a performance gain, achieved by minimizing required data transformations and –translations. This benefit, however, does not compete with the benefit offered by the loose coupling mechanism: IT flexibility. This is true for both the standardization of a component's *internal functionality* and for the standardization of it's *communication- interface*.

The essence of the standardization strategy is to remove any differences between components or parties. However, the idea that one can standardize everything (that one can have an agreement on everything) is obsolete. It is very questionable whether one should strive for a strong standardization like this.

Fortunately, data harmonization can also be achieved in the spirit of the loose coupling principle. For this purpose, the harmonization strategies A and B are combined into C: *standardization* of a central component instead of the decentral ones, and *bridging* between the decentral components on the one hand, and the central component on the other (see Picture 2, part *C*). In fact, loose coupling is nothing more than the bridging of differences between communicating components via a standardized decoupling component, specialized in harmonization.

Recommendations

1. Design your CDM to be suitable for all applications of data models.
2. Embrace the idea of loose coupling with conviction, and only standardize external components where it's feasible and useful.
3. Consider the CDM as one's own, organization wide model in which a data view on the organization is modeled, taking into account one's own culture and identity.
4. Make use of external standards as much as possible. Realize that their primary place is an external one, viewed from your organization.
5. External standards can easily be used as a source of ideas for the creation of a CDM. The motto here is: 'Better stolen well, than made up badly!' Please keep these points in mind:
   - Don't use external standards 1:1 to model or standardize organization-internal affairs, but only to borrow ideas from.
   - Whenever borrowing an idea, cut the dependancy immediately between the component that loans the idea and the component that borrows it (see Picture 1).
   - Be aware of the fact that many standards do not standardize on one dimension only, but on multiple dimensions simultaneously. Implementing such a standard can lead to inconsistencies with other standards or guidelines. Use standards that don't include multiple dimensions, or ones that allow you to separate dimensions easily. Only borrow ideas that fit into one dimension, without crossing dimension borders.
   - Model the semantics from your own vision on the data descriptions in the external standard, enabling you to identify and discuss possible interpretation differences with your business partners.
6. Become a member of all standardizing institutes relevant to your organization and offer an active contribution to the construction of standards. This way, you can have an influence making the standards fit your requirements as to external communication.

This article is a translated version of an earlier one written in Dutch, published in Computable, march 2007