CIDOC Documentation Standards Group
Correlation Test Project Initialization

January 26, 1999
By Martin Doerr

1. Introduction

Dear Friends and Colleagues,

Let me hereby initialize the Correlation Test Project for the object-oriented Conceptual Reference Model. This paper has the following structure: An introduction on the way of cooperation and communication, description of the objectives and approach, a work program and technical advice, all in the form of a proposal.

How do we start:
In the very first phase I'd like to achieve a general agreement on the overall goals and procedure. In order to realize a constructive and substantial discussion,
I shall present here a complete outline of the project, as my proposal and as questionnaire for any kind of additional or different opinions. Based on the feedback, I shall create a consolidated program, until all participants agree. All interested members should return this form filled in to me by mid of February.

How do we communicate:
First, we shall have a mailing list: cidoc-test@ics.forth.gr
This mailing list will contain any CIDOC members and their collaborators who want to join the project. It is an unmoderated list with controlled access. Anyone who wants to raise a decision can send a mail with the message "decision", a description of the decision point in yes or no form and a dead-line. The point is accepted, if either 50% of the list members answer positive, or at least 66% of the members answer and more than 50% of those are positive. All e-mails should be clearly marked in the subject area as "CRM Test". All messages all automatically logged.

Second, we shall have a Web site: http://www.ics.forth.gr/proj/isst/Activities/CIS/cidoc.
On this site, I shall keep:

- All documents of this project.
- A copy of the Preliminary Definition of the CIDOC Conceptual Reference Model, where we shall gather all entity-specific comments and proposals with reference to the creator and date of the comment.
- The descriptions of the individual teams participating and their tasks.
- The output of the work of the teams.
- A complete list of messages exchanged through the mailing list.
- A complete list of participants in the mailing list.

By that means, I try to create an efficient virtual discussion forum that allows anyone to understand the various opinions. Any further proposal is welcome, including groupware.
My idea is that this Web site is open to all CIDOC members. Alternatively, we may keep it between the (open list of) participants only, until a "representative form" is achieved, or
without any access control.
Third, we shall have two interim meetings in about 6 months distance, together with the
CIDOC data model group.

Decision point "How do we communicate":

[ ] I agree
[ ] I propose :

[ ] I do not understand:

2. Objectives:

1. Validate and improve the current contents of the CRM. As it is designed as an open, extensible standard, we would like to ensure, that it comprises the meaning of all concepts found in at least some of the real data formats used to describe material cultural objects. In particular we want to learn about the directions and needs for further development on the CRM.

2. Prove the utility of the CRM for specific applications. The CRM can help to improve the quality of mappings between database schemata, access profiles and data exchange formats. It can serve as unique intermediary between multiple schemata and multiple access profiles (views). It can help to generate compatible data exchange formats, as XML DTDs, new database schemata and access profiles. Finally it can help to harmonize our vocabulary and conceptualizations for data about material cultural objects, such that equivalent concepts can be more easily identified worldwide, and differences of non-equivalent concepts can be clearly formulated – with the ultimate goal of maximal interoperability between existing and future contents.

3. Gain experience in disseminating use of the CRM. The utility of the CRM depends critically on its comprehensive presentation. The group so far has made great attempts to present this model in three different forms already. These forms should be validated, eventually modified or others added. The textual parts, concept names, explanations and scope notes, should be improved. It should be tested, how obvious the relation between CRM concepts and other formats appears, and how the understanding of this relation can be improved.

4. Gain experience in mapping formats for material cultural object descriptions. The degree of deep semantic comparability between different actual descriptions and the degree to which these equivalences can be sufficiently formalized with the current means or those proposed here are an important question for any further attempt to mediate between cultural data. Equally important is the validation of mapping formalisms with respect to their wide comprehensiveness, ease of use and expressive power.

Decision point "Objectives":

[ ] I agree

[ ] I propose :

[ ] I do not understand:

[ ] For me has particular relevance:

3. Approach:

Tasks:
The test project preferably intends to coordinate related work already being done in established frameworks for the purpose of individual organizations rather than focus on new work specifically for CIDOC. Three

or four applications should be included to guarantee the appropriate generality and neutrality for a project at the CIDOC level. Several teams will volunteer to use the CRM in one or more of the following tasks:

- Identify the mapping from existing database schemata, access profile, metadata or other domain ontologies to the CRM.
- Formalize the mapping between existing database schemata, access profiles and metadata for retrieval.
- Reduce the number of mappings between multiple schemata and access profiles by using the CRM as an intermediary format.
- Create mappings between existing database schemata and document structures for data transfer.
- Create new document structures (DTDs) in XML format, database schemata, access profiles, metadata.

Immediate benefits for the participating teams are the exchange of knowledge, intellectual consistency between mappings from different teams, and the reuse of parts of mappings within or from other teams. In longer terms, these mappings will help implementers create "intelligent" access engines.

Decision point “Tasks”:

[ ] I agree
[ ] I propose :

[ ] I do not understand:

Procedure:
The whole procedure is envisaged for 1.5 years. The teams will declare the subject they intend to work on. In the sequence, they will begin to study the CRM and to compare it with their sources or the intended product (DTDs, schemata) in two feedback cycles. There will be a communication phase about the understanding of the CRM, which will result in an improvement of the comprehensiveness of its presentation, and an exchange of initial mappings or pieces of products, after about half a year. The latter will raise issues of conflicts or ambiguities in the sources of the teams, the CRM itself, and the mapping mechanism we propose. A phase of conflict resolution will follow, including a respective meeting. The result will be an improved mapping mechanism and new “working version” of the CRM for internal purpose of the project. Based on that, the final mappings or products are defined and a remaining list of open issues and conflicts identified. In a second meeting the formulation and packaging of all this experience in readable reports will be discussed, and until the end of the project, these reports will be created by interested members, not necessarily by all participants. A proposal to the Data Model Group will be made.

Decision point “Procedure”:

[ ] I agree
[ ] I propose :

[ ] I do not understand:

Outcome:
The expected outcome is a common public CIDOC report of all participants, which will describe this attempt to correlate data description concepts in the cultural area from a technical point of view. It will draw conclusions on this experience and make recommendations for further work in the field with respect to the development of the CRM, mapping of heterogeneous representations and the use of conceptual reference models (“domain ontologies”). It will contain as appendix the actual annotated mappings or new products as appropriate and a description of the final mapping formats used.

Further, a list of comments, additions and change proposals for each notion in the CRM and for its overall structure will stay available for further work on the CRM and other CIDOC activities.

Decision point “Outcome”:

[ ] I agree
[ ] I propose :

[ ] I do not understand:

5. Work Program:

I foresee the following steps:

1. Feed back from this paper, from the teams to the coordinator (me).
2. Fixing of the project program – consolidated proposal by the coordinator and feedback by the team.
3. Each team sends a short description of the task it has selected. For mappings, an annotated data dictionary of the selected schema or the definition of access points, etc. should be sent to be put on the project Web site. For new products, any other compact description of initial material, such as data standards etc. should be provided. (Please indicate clearly, if this raises any confidentiality concerns).
4. Start of the production phase: Communication cycle on the comprehension of the CRM and the mapping mechanism.
5. First input from the teams: Conflicts, ambiguities and initial pieces of products or mappings.
6. Conflict and ambiguity resolution. After respective electronic communication, a consolidated problem list will be the subject of a meeting, which will come up with a consolidated extension of the CRM and mapping method. The meeting will also decide, if this program has to be modified.
7. Second input from the teams (about 1 year after beginning): Final products, mappings, experience and conflict statements and recommendations.
8. Preparing the final report: A consolidated list of statements and recommendations and a layout for the final report is created by the coordinator (and/or any volunteer). A communication cycle is initiated on this layout and the assignment of authorship.
9. In the second meeting, the goals, style and conclusions from the project for the final report will be discussed, such that the individual authors agree on a common Leitmotif. The meeting will as well decide, if this program has to be modified.
10. Work on the report and exchange of draft report fragments.
11. Preparation of a final presentation for CIDOC 2000 ?
12. Final Presentation and decision for further work.

Decision point “Work Program”:

[ ] I agree
[ ] I propose:

[ ] I do not understand:

6. Technical advice

CRM:
With respect to the CRM, let me make these general remarks:

- The CRM approach is “maximal”, i.e., it should become as elaborate as the most specialized contents to be mediated. We therefore expect that each team will find that some concepts needed to do their work are missing in the current CRM. In particular, very general and very specific categories or attributes and complete thematic complexes may be missing.
- You are encouraged to extend the CRM to meet your specialized needs, assisted by members of the Group.
- You may uncover contradictions to current definitions in the CRM and unresolved methodological issues. Group members will cooperate to resolve such issues and to integrate the proposed extensions and modifications into the CRM. In particular the scope notes may need more detail, the explanations of the attributes and the relations between generic and specific attributes as well as the short cuts.

In any comment, please refer to all entities with the version date of the “Preliminary Definition of the oo CIDOC Reference Model”, the entity number and name. For instance:

P.D. Jan.25,1999 E11 Modification

For attributes, I propose a dot notation:

P.D. Jan.25,1999 E11 Modification.has produced (was produced by) : Man-Made Entity

Decision point “CRM”:

[ ] I agree
[ ] I propose:

[ ] I do not understand:

Mappings:

General idea:
Mappings are needed from each entity of the source to the appropriate entity of the CRM and from each attribute of the source to the appropriate attribute of the CRM. Each attribute combines two entities; attributes make no sense without the entities they combine. Therefore the entities must be mapped before the attributes can be mapped. In the easiest case, one attribute of your source will map to one attribute of the CRM. For instance, let's assume a Relational system, where a table for museum objects "musobj" has a field "material". This field contains keys of a controlled vocabulary implemented in the "matter" table, and is equivalent to the "consists of" in the CRM. Then you declare:

Table musobj	is exact equivalent to	P.D. Jan.25,1999 E19 Physical Object
Table matter	is exact equivalent to	P.D. Jan.25,1999 E59 Material
Field musobj.material	is exact equivalent to	P.D. Jan.25,1999 E18 Physical Entity. consists of (is incorporated in): Material

The CRM does not deal with the implementation of keys. Therefore the equivalence with a Relational system must be seen on the level of an Entity Relationship diagram. In particular any key in a Relational system should map to an entity in the CRM. The CRM is object-oriented. Therefore an attribute may be declared in a superclass of the mapped entity. In the above example, the attribute "consists of (is incoporated in)" is declared for Physical Entity and therefore holds for Physical Object as well (because Physical Entity is superclass of Physical Object).
Obviously the mapping can only be understood, if the annotated data dictionary of the respective source is available. In the case of a new product, the mapping can be from the CRM to the product or vice-versa.

Decision point "General idea":

[ ] I agree
[ ] I propose:

[ ] I do not understand:

Mapping Entities:
Let us regard entities as notions that group things together by common explicit or implicit characteristics, like concepts in a thesaurus. They refer to some set of items in reality, their "extension" or "instances". The superclass relation in the CRM corresponds to the notion of a broader term, i.e. it comprises all potential instances of all its narrower terms/ subclasses. If we compare a system of entities to those of the CRM, I propose to apply the extended expressions of ISO5964 as in the Getty Guide-Lines for Forming Language Equivalents:

1. Exact equivalence
2. Broader equivalence
3. Narrower equivalence
4. Inexact equivalence.
5. No equivalence.

In case 3 and 4 an explanation is needed, which instances of the source are not covered by the respective CRM entity and why. Maybe a rule must be given as to which values the mapping does not hold for. One source entity may map to more than one CRM entity. Again in this case a rule for the distribution of the instances should be supplied.

E.g., let us assume the instances of the table "musobj" may map to "E24 Iconographic Object" only, if the classification term stored in "musobj.objname" is any narrower term of AAT "figural works".
Instances that cannot be mapped and differences that cannot be resolved by rules are regarded as mapping conflicts, whereas a plain "no equivalence" is regarded as an omission of the CRM, a case for extension.

There can be hidden entities in a source, flags and encodings that hide references to meaningful entities in a primitive data type as integer or string.

Decision point “Mapping Entities”:

[ ] I agree
[ ] I propose:

[ ] I do not understand:

Mapping Attributes:

Let us regard all attributes as roles. An attribute connects a base entity (typically shown on the left side) with a “value”, another entity (typically shown on the right side). The kind or name of the attribute declares the role the attribute value plays for the base entity. E.g. let Martin be a person and 47 a number, then Martin.age = 47 says that 47 plays the role of “age” for the person Martin. Under this view differences between attributes and references are merely an implementation detail. Structures within structures, e.g. the inclusion of fields for an address in the record for a person, can be dealt with as a reference to such a structure, and this reference as a simple attribute to an entity “address”, which is in turn analyzed into its fields. This holds in particular for DTDs which describe mostly nested structures.

When comparing attributes, we may again encounter differences of the scope we may express by one of the five equivalence criteria: exact, broader, narrower, inexact, none, as discussed for entities. In addition, one attribute in one model may compare only to a path of subsequent attributes in another model:

Field musobj.creator is exact equivalent to

P.D. Jan.25,1999 E18 Man-Made Entity.was produced by (has produced): Modification-
E7 Activity.carried out by (performed): Actor

Actually we have referred here to the attribute "Modification.has produced (was produced by): Man-Made Entity" in its inverse form for the obvious reason to maintain the direction of the mapped attribute "creator". All the CRM attributes have been designed symmetrically for this reason. In some cases, the source model may be more detailed, such that a path of the source is needed to map to a CRM attribute.

I deliberately do not give more details here for the clarity of the basic issue. I expect that the above mechanism will cover some 90 per cent of the cases. I'd encourage people first to make a practical trial with this form, until we go into more details. Of course the use of graphics can make attribute mappings far more comprehensible and is encouraged. We may select some good presentation as template for further work.

Decision point "Mapping Attributes":

[ ] I agree
[ ] I propose:

[ ] I do not understand: