DCPAPERS

Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 139-146

Using Dublin Core to Build a Common Data Architecture

Sandra Fricker Hostetter
Rohm and Haas Company, Knowledge Center
shostetter@rohmhaas.com

Abstract

The corporate world is drowning in disparate data. Data elements, field names, column names, row names, labels, metatags, etc. seem to reproduce at whim. Librarians have been battling data disparity for over a century with tools like controlled vocabularies and classification schemes. Data Administrators have been waging their own war using data dictionaries and naming conventions. Both camps have had limited success. A common data architecture bridges the gap between the worlds of tabular (structured) and non-tabular (unstructured) data to provide a total solution and clear understanding of all data. Using the Dublin Core Metadata Element Set Version 1.1 and its Information Resource concept as building blocks, the Rohm and Haas Company Knowledge Center has created a common data architecture for use in the implementation of an electronic document management system (EDMS). This platform independent framework, when fully implemented, will provide the ability to create specific subsets of enterprise data on demand, enable interoperability with other internal or external systems, and reduce cycle time when migrating to the next generation tool.
Keywords: *common data architecture, CDA, document management, platform independent framework, data resource management, metadata, Dublin Core, controlled vocabularies*

1. A new hybrid

Organizing information has become a core competency for corporations. Moving from a paper-based world to an electronic-based one is a difficult and lengthy transformation. Paper forced us to behave in certain ways because of physical limitations associated with its tangibility. However, paper also had inherent strengths in its universality and this is something we have taken for granted.

Blending the features of paper and electronic formats is an enormous challenge. We must create something new. The plant world provides us with a helpful analogy. A hybrid plant is the combination of two separate entities into something completely new and unique, yet shares the attributes of both parent plants. This does not happen by accident. Two different species of plants will not merge to create a new one without purposeful human intervention, management, and care. And therein lie both the problem and the opportunity.

In the past, tabular and non-tabular data have been managed and accessed in very different ways. However, the ever-demanding user population wants to see all the available data integrated together and presented in a manner individually tailored to their specific needs. It has become impossible to separately manage non-tabular data and tabular data. This demands we address seemingly mutually exclusive issues in a way that satisfies all parties. The creation of a common data architecture is the most effective way to bridge the gap between all types of data.

2. Metadata management in a document managed world

The importance of controlling the metadata used to describe items deposited in a document management system is critical to facilitate effective search and retrieval activities in partnership with the dueling aspects of a full-text environment – instant gratification and lack of discrimination. At the Rohm and Haas Company, Dublin Core was a good starting point and became the basis for the document class and document properties structure "dictated" by the EDMS. From the beginning, our goal was to create a platform independent framework that would meet the following needs: (1) enable the creation of specific subsets of enterprise data on demand (2) provide future interoperability with other internal and external systems (3) reduce cycle time when migrating from "today's tool," to the next generation of document management software without excessive rework.

https://doi.org/10.23106/dcmi.952107045

The Dublin Core data elements as implemented in the EDMS at the Rohm and Haas Company function as the common metadata. All document classes have these properties, though it is not mandatory the properties be populated. Eventually, three of these Dublin Core based properties (DC.Title, DC.Date.issued, DC.Publisher) will be required, and DC.Publisher will have a Rohm and Haas specific controlled scheme to reflect the company's business unit structure.

3. The common data architecture approach

A common data architecture (CDA) "is a formal, comprehensive, data architecture that provides a common context within which ALL DATA are understood and integrated". A CDA has the following basic components – data subjects, data characteristics, and data characteristic variations. A *data subject* is "a person, place, thing, concept, or event that is of interest to the organization and about which data are captured and maintained". A *data characteristic* is "an individual characteristic that describes a data subject". A *data characteristic variation* "represents a difference in the format, content, or meaning of a specific data characteristic" (Brackett, 1994, p. 31, p. 39).

At first glance, a standard like the Dublin Core Metadata Element Set Version 1.1 looks like it might be a common data architecture. However under closer scrutiny, its deficiencies become more obvious. Dublin Core violates a core principle of data management by mixing different facts within a single field. DC.Creator can represent a person or an organization. The ideal data management equation is 1 Fact = 1 Field. In Dublin Core's well-intended effort to be simple yet fully extensible, it is also very non-specific. This leads us down the tempting path to the never-ending crosswalk. Cross walking happens only at the physical level, requires an excessive amount of work, and yields minimal understanding. Instead, if we move beyond the traditional physical level analysis and cross-reference to a common data architecture created at the logical level, we gain a true common context for understanding all data.

4. How to build a common data architecture

Building a common data architecture involves five major steps. It is a reiterative process that may take several months to become an accurate reflection of the organizational situation and will require occasional readjustments over time. Since a common data architecture represents is a living breathing organization that grows and changes, it too must be refreshed as needed.

4.1 Defining the "pivotal" data subject

The first step is to identify, formally name, and define the pivotal data subject. The pivotal data subject is the most central business concept. All related concepts will be organized around this data subject. The pivotal data subject for the EDMS was the software defined object "Document Class". We adopted the Dublin Core terminology for "Information Resource" and broadened the definition as follows:

Information Resource

An Information Resource is a set of data in context, recorded in any medium of expression (text, audio, video, graphic, digital) that is meaningful, relevant, and understandable to one or more people at a point in time or for a period of time. Traditionally, an Information Resource is recorded on some medium, such as a document, a web page, a diagram, and so on. In the broad sense, however, an Information Resource could be a person or a team of people.

An Information Resource in this data architecture represents a version of an Information Resource when there is more than one version produced. The Information Resource. System Identifier changes for each version. The Information Resource Document. Number that is assigned as an Information Property Item through Information Resource Property remains the same across versions and identifies the Information Resource, and the Information Resource. Version Identifier uniquely identifies the version of that Information Resource.

Note that the system identifier as defined in this data architecture is the system identifier of the home system where data about information resources are stored. Any other foreign identifiers from other systems where data about information resources are stored are assigned as an Information Property Item through Information Resource Property.

Note that there are non-EDMS versions of an Information Resource, such as web page versions, that may not have a date, version identifier, URL change, and so on. There is no way to know or distinguish versions of this type.

4.2 Defining the data characteristics

The second step is to identify, formally name, and define the data characteristics of the pivotal data subject. Examples include:

Information Resource. Title

The official title of the Information Resource, such as "The Importance of Adding Property Data to a Panagon Document." This is the name by which the Information Resource is formally known.

https://doi.org/10.23106/dcmi.952107045

Information Resource. System Identifier
The system assigned identifier in the home system that uniquely identifies an Information Resource. This is not the same as the system identifier that identifies an Information Resource in an EDMS system or any other foreign system documenting Information Resources. The Information Resource, System Identifier changes for each version of an Information Resource. The Information Resource. Version Identifier identifies the version of the Information Resource.

Information Resource. Version Identifier
The version number of the Information Resource. The versions are typically, but not necessarily, assigned sequentially from 1. In some foreign systems or standards, the version identifier may be appended to the system identifier. In this data architecture, the version identifier is maintained separate from the system identifier.

Information Resource Subtype. Code
Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. Not every Information Resource Type will have Information Resource Subtypes.

Information Resource Type. Code
The code that unqiuely identifies an Information Resource Type, such as LNBK for the Information Resource Type Laboratory Notebook.

4.3 Defining the qualifying data subjects

The third step is to identify, formally name, and define any qualifying data subjects and their data characteristics. We used the Dublin Core Metadata Element Set Version 1.1 as the basic building blocks. Examples include:

Information Contributor
An Information Contributor is any person or organization that contributes in any way to an Information Resource. A person may be an author, a researcher that provides material, or a reviewer, and an organization may be a service or professional organization. Information Resource Contributor connects an Information Contributor to an Information Resource. Information Resource Contributor Role identifies the specific role played by an Information Contributor.

Information Property Group
An Information Property Group is a set of related Information Property Items. The structure of Information Property Groups and Information Property Items allows a variety of reference tables or enumerated lists to be defined for assignment to an Information Resource through Information Resource Property. Information Property Group represents a controlled set of reference tables

Information Property Item
Information Property Item is one reference item from a set of reference items commonly held by an Information Resource. Each Information Property Item belongs to an Information Property Group. Information Resource Property assigns the Information Property Items to Information Resources.

Information Property Item Alias
An Information Property Item can have different names in different systems or standards. There is no uniform name that transcends all systems and standards. Information Property Item Alias documents all of the alias names for a foreign Information Property Items in various systems and standards, and their originating system or standard. The preferred name is shown in Information Property Item. Name.

Information Resource Contributor
An Information Resource can have many different Information Contributors, and an Information Contributor can contribute to many different Information Resources. Information Resource Contributor designates a specific Information Contributor for a specific Information Resource. Information Resource Contributor Role identifies the specific role performed by an Information Resource Contributor.

Information Resource Contributor Role
An Information Contributor can perform different roles with respect to an Information Resource. Information Resource Contributor Role is a reference table identifying the roles that an Information Contributor can perform for an Information Resource.

Information Resource Property
An Information Resource can be characterized by many different Information Property Items, and an Information Property Item can characterize many different Information Resources. Information Resource Property assigns a specific Information Property Item to a specific Information Resource. If that Information Property Item requires additional data, such as a date or description, those data are provided in the data characteristics described below.

Information Resource Property Validity
An Information Resource Type has a set of Information Properties Items that are valid and can be assigned to an Information Resource belonging to that Information Resource Type. Information Resource Property Validity indicates the valid assignments of Information Property Items. Note that this data subject is set up to show only the valid assignments of an Information Property Item for an Information Resource Type. If an Information Property Item appears, then that Information Property Item is valid for the Information Resource Type. If an Information Property Item does not appear, then that Information Property Item is not valid for the Information Resource Type.

https://doi.org/10.23106/dcmi.952107045

Information Resource Publisher

An Information Resource can be published by more than one Publisher, and a Publisher can publish more than one Information Resource. Information Resource Publisher identifies the publication of an Information Resource by a specific Publisher.

Information Resource Relationship

An Information Resource can have a relationship with other Information Resources, such as reference to another Information Resource, material included from another Information Resource, and so on. Information Resource Relationship identifies a specific relationship between two Information Resources. Information Resource Relationship Type identifies the specific type of relationship between Information Resources.

Information Resource Relationship Type

Information Resource Relationship Type is a reference table that identifies the specific type of relationship between two Information Resources identified in Information Resource Relationship.

Information Resource Subtype

Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. Not every Information Resource Type will have Information Resource Subtypes.

Information Resource Type

Information Resource Type is a broad grouping of Information Resources that designates the nature or genre of the content of the Information Resource. It describes general categories, functions, or aggregation levels of the content of Information Resources.

Information Security Group

An Information Resource can have different levels of security classification governing which individuals or organizations can access that Information Resource. Information Security Group is a reference table designating the broad levels of security for an Information Resource. Information Security Subgroup identifies a more detailed grouping of security.

Information Security Subgroup

Information Security Groups can have a more detailed level of classification. Information Security Subgroup provides the detailed levels of security classification within Information Security Group.

Publisher

A Publisher is any organization, internal or external to Rohm and Haas, that formally publishes an Information Resource. Note that this current definition is limited to Information Resources. As the common data architecture is enhanced, this definition may be altered to include the publishers of other material not considered an Information Resource.

4.4 Creating a visual representation of the relationships

The fourth step is to create a visual representation of how all the data subjects relate to each other. In Figure 1 the relationships are depicted in a manner based on data modeling techniques outlined below:

Arrows moving away from a data subject represent a one-to-many relationship between the data subjects. For example, a single Information Resource may have many Information Resource Contributors. An Information Contributor (DC.Creator or DC.Contributor) is any person or organization that contributes in any way to an Information Resource. A person may be an author, a researcher that provides material, or a reviewer, and an organization may be a service or a professional organization

Arrows moving towards a data subject represent a many-to-one relationship. For example, an Information Contributor may be an Information Resource Contributor to many different Information Resources. Information Resource Contributor designates a specific Information Contributor for a specific Information Resource. Information Resource Contributor Role identifies the specific role performed by an Information Resource Contributor.

Two arrows represent a relationship between two Information Resources. For example, an Information Resource can have a relationship with other Information Resources, such as reference to another Information Resource, material included from another Information Resource, etc. Information Resource Relationship identifies a specific relationship between two Information Resources. Information Resource Relationship Type identifies the specific type of relationship between Information Resources.

Multiple arrows going in the same direction in sequence represent a hierarchy relationship. For example, Information Resource Subtype indicates a more detailed classification of documents within Information Resource Type. However, not every Information Resource Type will have Information Resource Subtypes.

Arrows going towards each other and intersecting at the same data subject represent an assignment relationship. For example, Information Resource Contributor connects an Information Contributor to an Information Resource. Information Resource Contributor Role identifies the specific role played by an Information Contributor (the various Information Resource Contributor Roles that an Information Contributor can perform for an Information Resource are stored in a reference table. This will be discussed in more detail under heading 5. Properties Make the World Go 'Round). We assign an Information Contributor to an Information Resource and then we give the Information Contributor a specific role. The same kind of assignment relationship exists for Publisher. An Information Resource could be published by two different publishers. The print copy could be pub-

https://doi.org/10.23106/dcmi.952107045

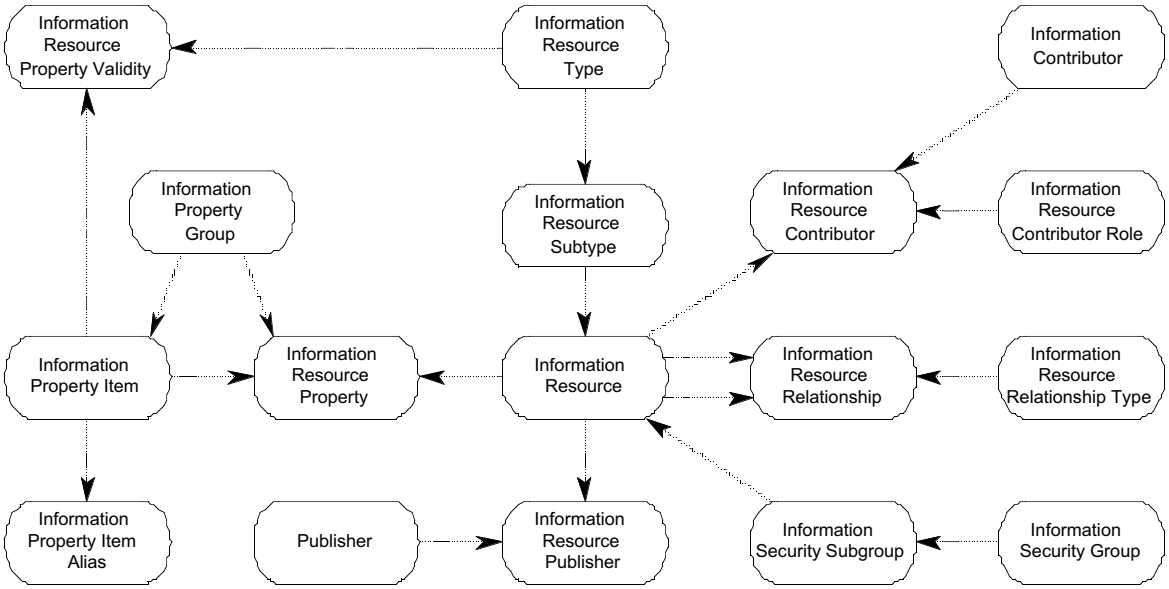

Figure 1.

lished by a different entity than the electronic version and the electronic and print versions could be the same content or might be different content.

4.5 Testing and adjustment

The final step is to test the resulting common data architecture and adjust as needed. This can be done by trying it out on another system or conducting business use cases.

5. Properties make the data world go 'round

Properties (fields, attributes, characteristics, features, metatags) help us understand more about the content and context of the information resource to which they belong. Common properties are universal. Everyone in the organization cares about these properties. It is important to limit the names and display labels of these common properties so we can effectively share them and mean the same thing. Special or custom properties apply only to a small subset of information resources, but their names and labels should be limited also. Limiting the values for most properties helps keep the context meaningful and clear.

Because an Information Resource may have many different Information Resource Property Items, we need to resolve the many-to-many relationship and figure out a way to assign them to the specific Information Resource. We define the Information Resource Property Items first, and then assign them. By structuring things in this manner, Information Property Groups and Information Property Items within those groups can become ineffective at any time without altering the structure of the data resource (Figure 2).

Information Resource Property Items for a specific Information Resource are kept in reference tables called Information Property Groups. An Information Resource Property is a qualifying Data Subject which assigns Information Resource Property Items, via the Information Resource Property Groups structure to a specific Information Resource.

Information Property Group is a reference table of reference tables. Information Property Item is a specific value in a reference table. All Information Property Items must belong to an Information Property Group. This portion of the CDA represents "a controlled vocabulary of controlled vocabularies". These reference tables are documented as data subjects, but their definitions clearly identify them as reference tables and not true data subjects.

An example of an Information Resource Property

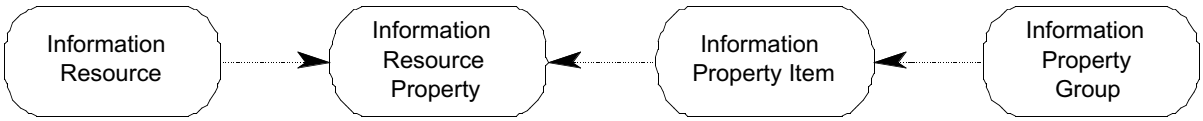

Figure 2.

https://doi.org/10.23106/dcmi.952107045

Group is Information Resource Description. An Information Resource can have many associated descriptions, such as content, spatial, physical format, temporal, and rights. Information Resource Description identifies each of the descriptions that can be assigned to an Information Resource. Examples of Information Property Items for the Information Resource Property Group called "Information Resource Description" are Content Description, Spatial Description, Physical Format Description, Temporal Description, and Rights Description. Other examples of Information Property Groups include: Information Resource Date, Information Resource Identifier, Information Resource Library, Information Resource Subject, Language, and Non-Enumerated Feature.

6. Documenting the common data architecture

We are formally documenting our common data architecture in the Data Resource Guide. The Data Resource Guide is a proprietary Microsoft Access software application which contains tables on the Common Data Architecture side for data subject, data characteristic, data characteristic variation, data code set, data code. On the Data Product side (e.g. EDMS, Dublin Core Metadata Element Set Version 1.1, etc.) the database has tables for data product type, data product, data product group, data product unit, and data product code. It also has tables for data product cross-referencing. The inclusion of a reporting feature enables the data resource administrator to see how multiple data products relate to each other and what data elements they share.

7. Cross referencing Dublin Core to the CDA

When we cross-reference the Dublin Core Metadata Element Set Version 1.1 to the common data architecture it yields the following results.

Dublin Core Element Label	Common Data Architecture Equivalent (Data Subject. Data Characteristic, Data Characteristic Variation)
Title	**Information Resource. Title, Variable**
Creator	Creator can be either a person or an organization. The cross-references are identified for each Creator variant. **Information Contributor. Person Name, Complete Inverted** Comment: Information Resource Contributor Role. Name, Formal = 'Creator' **Information Contributor. Organization Name, Variable** Comment: Information Resource Contributor Role. Name, Formal='Creator'
Subject	An exact cross-reference is indeterminate based on the definition of Subject and the lack of a specific controlled vocabulary or formal classification scheme. Any implementation could use one or more controlled vocabularies or formal classification schemes. The best cross-reference approach is to identify each specific controlled vocabulary or formal classification scheme used under the Dublin Core standard, document it as a reference table in the common data architecture, and then prepare a cross-reference to that reference table. **Business Unit Classification Scheme. Name, Formal** Comment: Information Property Group. Name, Formal = Information Resource Subject Comment: Information Property Group. Name, Formal is indeterminate and needs to be determined for each data occurrence.
Description	Description is defined as a reference table in the common data architecture as Information Resource Description. The specific types of descriptions, such as table of contents, abstract, etc. are reference items in that reference table. **Information Resource Property. Description, Dublin Core** Comment: Information Property Group = Information Resource Description Comment: Information Property Item. Name, Formal is variable and needs to be determined for each data occurrence.
Publisher	**Publisher. Name, Variable** Comment: The publisher name should be used as the cross-reference.
Contributor	Contributor can be either a person or an organization. The cross-references are identified for each Contributor variant. **Information Contributor. Person Name, Complete Inverted** Comment: Information Resource Contributor Role. Name, Formal is variable and needs to be determined for each data occurrence. **Information Contributor. Organization Name, Variable** Comment: Information Resource Contributor Role. Name, Formal is variable and needs to be determined for each data occurrence.
Date	Date is defined as a reference table in the common data architecture as Information Resource Date. The specific types of dates, such as Available Date, Creation Date, Issued Date, Modified Date, Valid Date, etc. are reference items in that reference table. **Information Resource Property. Date, ISO 8601**

https://doi.org/10.23106/dcmi.952107045

	Comment: Information Property Group. Name, Formal=Information Resource Date Comment: Information Property Item. Names, Formal is variable and needs to be determined for each data occurrence.
Type	**Information Resource Type, Name, Dublin Core** Comment: If a controlled vocabulary other than the list of Dublin Core types is used, it needs to be documented as a data product unit variant and cross-referenced to an appropriate reference table in the common data architecture.
Format	Format is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of descriptions, such as text, audio, etc. are reference items in that reference table. **Information Resource Property. Description, Dublin Core** Comment: Information Property Group. Name, Formal = Information Resource Description Comment: Information Resource Description. Name, Formal is variable and needs to be determined for each data occurrence.
Identifier	Identifier is defined as a reference table in the common data architecture as Information Resource Identifier. The specific types of identifiers, such as URI, ISBN, etc., are reference items in that reference table. **Information Resource Property. Value, Variable** Comment: Information Resource Property = Information Resource Identifier Comment: Resource Description. Name, Formal is variable and needs to be determined for each data occurrence
Source	Source represents a relationship between two Information Resources as defined in Information Resource Relationship. The identifier of the Source in Dublin Core must be determined, the system identifier located, and that system identifier used in Information Resource Relationship. System Identifier. The specific types of relationships, such as source, and so on, are defined in Information Resource Reference Type. **Information Resource Property. Value, Identifier Variable** Comment: Information Resource Relationship Type. Name, Formal = Source
Language	Language is a multiple-fact data item for the language and the country associated with the language. The cross-references are identified for each language variant. **Language. Code, ISO 639** **Country. Code, ISO 3166**
Relation	Relation represents a relationship between two Information Resources as defined in Information Resource Relationship. The identifier of the Source in Dublin Core must be determined, the system identifier located, and that system identifier used in Information Resource Relationship. System Identifier. The specific types of relationships, such as source, etc. are defined in Information Resource Reference Type. **Information Resource Property. Value, Identifier Variable** Comment: The Information Resource Relationship Type. Name, Formal is indeterminate and needs to be identified for each data occurrence.
Coverage	Coverage is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of coverage, such as spatial, temporal, etc. are reference items in that reference table. Information Resource Description. Name is variable and needs to be determined for each data occurrence. **Information Resource Property. Description, Dublin Core** Comment: Information Resource Property. Name, Formal = Description Comment: Information Property Item. Name, Formal = Spatial Description Comment: Information Property Item. Name, Formal = Temporal Description Comment: Information Property Item. Name, Formal = Jurisdiction Description
Rights	Rights is a specific type of description which is defined as a reference table in the common data architecture as Information Resource Description. The specific types of rights, such as copyright, royalty, and so on, are reference items in that reference table. **Information Resource Property. Description, Dublin Core** Comment: Information Property Group. Name, Formal = Information Resource Description Comment: Information Property Item. Name, Formal is variable and needs to be determined for each data occurrence.

8. Next steps

This common data architecture is currently a work-in-progress. Full documentation of the common data architecture in a Data Resource Guide must be completed as well as the final cross-referencing of the EDMS metadata and Dublin Core Metadata Element Set 1.1. The creation of a thesaurus component is essential to making the CDA content available to the wider community of general

https://doi.org/10.23106/dcmi.952107045

system users and the individuals who develop and design new database applications. We envision a "Data Element Supermarket" where developers can shop for the field name desired, find its variations (code, name, acronym), and learn its single source and history of use in other systems. We have created a good foundation, but there is still much work to be done before the true value can be realized.

Acknowledgements

This work would not have been possible without the professional advice and consultation expertise of Michael H. Brackett. The author is grateful for his personal encouragement and support.

References

1. Brackett, M., 1994. *Data Sharing Using a Common Data Architecture*. New York: John Wiley & Sons, Inc.

https://doi.org/10.23106/dcmi.952107045