1.09 Objavljeni strokovni prispevek na konferenci
1.09 Published Professional Conference Contribution

Bogdan Florin Popovici[*]

SOME CONCEPTUAL DATA MODELS FOR RECORDS AND ARCHIVES

Abstract:

ISAD(G) is, by far, the most acknowledged standard for archives, that also comprise a data model, about the structure and relation between archival entities. Besides, there are also some other attempts to codify these complex relationships and also those of creators or functions. Also, there have been developed some data models for records management that might be interesting to be regarded and connected to archival models, as part of the same lifecycle of records. The present paper aims to give a brief presentation of these data models, with some critical remarks.

Key words:

archival data models, archival entity, MoReq, CIDOC-CRM PSI, PREMIS, The Australian Recordkeeping Metadata Schema, ICA descriptive standards, Hurley's Common Practice Rules

Izvleček:

Nekaj konceptualnih podatkovnih modelov za dokumentarno in arhivsko gradivo

ISAD(G) je zdaleč najbolj priznan standard za arhivsko gradivo, ki prav tako sestavlja podatkovni model, o strukturi in relaciji med arhivskimi entitetami. Obstaja tudi nekaj drugih poskusov kodiranja teh kompleksnih relacij ter tudi relacij ustvarjalcev in funkcij. Razviti so bili tudi nekateri podatkovni modeli za upravljanje z dokumenti, ki bi lahko bili zanimivi za povezovanje z arhivskimi modeli kot del istega življenjskega cikla dokumentov. Prispevek poskuša, z nekaj kritičnimi pripombami, podati kratko predstavitev teh podatkovnih modelov.

Ključne besede:

arhivski podatkovni modeli, arhivska entiteta, MoReq, CIDOM-CRM PSI, PREMIS, avstralska metapodatkovna shema, standardi za popisovanje MAS, Hurleyjeva splošna pravila

For a long time, records managers and archivists minded their own business in dark and dusty rooms (or, at least, this is how the others looked upon them…). Their activity was essentially considered as a practical one, the folder was a folder, the record was a record - no rocket science.

The advent of Information Technology brought several challenges, among which:

- how can an archivist define her/his working framework in such a way that the workflow is automatized and the PC may work for her/him (for instance, "what descriptive elements do I need as the online finding aid is fully understandable by user?")

[*] *Bogdan Florin Popovici, Ph. D., archivist, Arhivele Nationale, Str. Gh. Baritiu nr. 34, 500025 Brasov, Romania, email: bogdanpopovici@gmail.com.*

- what is the abstract nature of her/his work object, that is the same whether the record is on paper or in a digital form and that shapes her/his working methods and processing approach (for instance, "which are the common features of a medieval charter and a You Tube record?")

In this regard, many archival scientists and practitioners delivered detailed answers, produced standards, identified "proper" metadata up to the extent when one needs new finding aids in order to control the number of different standards. The goal of this paper is to have a brief review of several conceptual data models for records and archives, in order to notice various approaches about the archival entities[1].

A. MoReq Data Model[2]

MoReq was developed in the beginning of 2001, as a specification for European electronic records management systems. The Specification has, until now, 3 versions (2001, 2008 and 2011, respectively MoReq, MoReq2 and MoReq2010). The second version marked a climax of a certain type of approach that will eventually be changed with MoReq2010. It is not the place or time to ponder between pros and cons of each version. I would only point out that MoReq2 is more based on classical records management and its conceptual model is illustrative in this regard. On the other hand, MoReq2010 is more IT oriented, although it has a strong, abstract and modern representation of electronic records management principles; it does not have a specific data model for archival entities, but I shall try to emphasize its peculiarities in comparison with MoReq2, for the purpose of this paper.

A1. MoReq2

The data model envisaged by MoReq2 is represented in the Specification in the *Figure 13.3* and reproduced here as appendix 1. In this model, the component is the smallest individual part; it is an element for the creation of records or of documents that make a record. Both records and documents might have different types. Records are elements of files (with their subdivision, sub-files or volumes) and also of classes (that derives from a classification scheme). A records and disposition schedule applies to all entities (class, file and its divisions and record type).

A2. Moreq2010

As I mentioned before, MoReq2010 does not have an explicit entity model. As a refactored product of MoReq2, it might be assumed it follows, in main lines, the model of MoReq2. But MoReq2010 goes further, using the power of IT systems for a modern electronic records management (see appendix 2). In this regard, Moreq2010 assumes a new element, below component: the *content* (Fig. 6c). Although not

---

[1] *In this paper, the term "archives" or "archival" will not have the usual meaning in US or UK tradition, but it will encompass the whole lifecycle of records, ether of primary or secondary value; that is, records and archives will be treated as a whole, as groups of records. Also, as a general remark, all the figures, diagrams etc. in the appendix of this paper are reproduced from the original works cited at the beginning of each section and they are inserted for illustrative purpose only. All the rights for these figurative appendixes rest on the copyright holders of the original sources.*

[2] *European Commission, Model Requirements For The Management Of Electronic Records Update And Extension, Moreq2 Specification, Brussels, 2008; idem, MoReq2010® Modular Requirements for Records Systems, Volume 1: Core Services & Plug-in Modules, Version 1, Brussels, 2011.*

explicitly defined in the text, the content is the ultimate target for management and preservation: in the last instance, the core element in a class, in a folder, in a record, in a document is the content. Due to high versatility of electronic documents, content might be a record or might be a document, and it is important to manage it with the proper tool for records management.

Furthermore, although MoReq2010 keeps supporting the classical classification scheme (Fig. 201b), it introduces the possibility of having multiple classification and different aggregation of records (Fig 201c). That is, a record can be classified in one class for aggregation, in another class for content and in another class for retention. That creates new, multiple relationships between records and groups of records.

Therefore, the data model that results from MoReq2010 specification, for electronic records seems to outline a data model where records are made of components and content, and they may be part of one or many classification and/or aggregation. It should be noticed the carrier/medium is not relevant in this model.

B. Public Sector Information According to CIDOC-CRM[3]

CIDOC-CRM is a conceptual data model, developed initially within the International Committee for Documentation of the International Council of Museums. The model was elaborated between 2000 and 2006 and it was accepted as ISO 21127 standard. The model provides ontology for concepts and information in cultural heritage[4].

The modelling of public sector information (i.e. records), presented graphically in appendix 3, is based on 3 main entities: *public administration* (a legal body) that performs a *function* (seen as an activity) in the presence of a *record* (as a man-made object). Each of these has different other sub-entities and they relate to each other under the proposed connections.

Beyond the conceptual strictness of the model, one can argue that many elements presented there are not very relevant for records or archives management (for example, *life event*) or, in other cases, essential elements are not comprised (for instance, *date* for a record)

C. PREMIS[5]

Preservation Metadata: Implementation Strategies was developed with the intent of identifying the proper metadata to support digital preservation activities. After the initial project, between 2003 and 2005, currently PREMIS is maintained by the Library of Congress.

Although not specifically developed for archives but as a standard for preservation metadata, PREMIS is used in modelling archival information systems[6]. PREMIS proposed a data model, as seen in appendix 4. In one scenario of

---

[3] *Lina Bountouri, Christos Papatheodorou and Manolis Gergatsoulis, Modelling the Public Sector Information through CIDOC Conceptual Reference Model at http://www.cidoc-crm.org/docs/64280404.pdf (last visit 15. 11. 2012).*

[4] *Information retrieved from http://www.cidoc-crm.org/ (last visit 15. 11. 2012).*

[5] *http://www.loc.gov/standards/premis/.*

[6] *This information resulted in different discussions I had with 2-3 suppliers of AIS.*

interpretation, an Object (a *record*, for instance) was generated by an Agent (that has certain Rights over that object), during an Event. The component might be part of an Intellectual Entity (a *file*, for instance).

D. The Australian Recordkeeping Metadata Schema (RKMS)[7]

The RKMS was adopted in 1999 and is based on the holistic approach of records continuum. It outlines, among others, *"the connections between* business, *defined broadly to encompass all social and organisational activity, the* people or agents *who do business, and* the records *which are by-products of that business*"[8]. This model of interactions has been included in an international standard[9].

In this model, the entities connected with the creation of Records are Agents (People) and Business. The Agents *do* Business that is *documented* in Records. These last ones are created, managed and used by Agents. In a larger context, it is identified a Mandate, that govern the Business, establish competencies of Agents and that is accounted for its execution by Records.

ICA DESCRIPTIVE STANDARDS[10]

In the last decades, International Council of Archives undertook serious efforts towards the standardisation of archival description. The outputs were 3 standards for describing archival materials (ISAD-G), creators (ISAAR-CPF) and functions (ISDF)[11].

Based on the ISDF (see Appendix 5 of the present paper), a function is performed by a corporate body and it is documented in records. Records are created, managed and used by corporate bodies. Each entity involved presents a hierarchical structure of sub-entities. For archival material, documents, files, series, sub-fonds and fonds[12] can be identified (see appendix 7); for creators, there might be subdivisions of a corporate body (for example, minister, directorate, division, service, office etc.); for functions, it might be also a hierarchy, from mandate, business, function, activity, transaction etc.

---

[7] *Glenda Acland, Barbara Reed, Sue McKemmish, Documenting Business: The Australian Recordkeeping Metadata Schema at http://www.infotech.monash.edu.au/research/groups/rcrg/publications/adcs.html; see also http://www.infotech.monash.edu.au/research/groups/rcrg/projects/spirt/deliverables/rkmsgen-tech-intro.pdf.*

[8] *Acland, Reed, McKemmish, loc. cit.*

[9] *ISO 23081-1:2006, Information and documentation —Records management processes —Metadata for records — Part 1:Principles.*

[10] *See the following resources: ICA-CBPS, Progress report for revising and harmonising ICA descriptive standards, 2012 (http://www.ica.org/13155/standards/cbps-progress-report-for-revising-and-harmonising-ica-descriptive-standards.html); ICA - CBPS, Relationship in archival descriptive systems, 2012 (http://www.ica.org/13149/standards/cbps-relationship-in-archival-descriptive-systems.html), ISAD(G) (http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html); ISAAR-CPF (http://www. ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html); ISDF (http://www.ica.org/10208/standards/isdf-international-standard-for-describing-functions.html).*

[11] *ISDIAH, as a standardized description of archives holders, is an instance of ISAAR-CPF and I do not consider it relevant for the purpose of this paper.*

[12] *Here might be a broader discussion, if all sub-entities are mandatory or not, if the enumeration is complete or not. This kind of analysis goes over the intents of present paper.*

ICA Atom[13]

Although based on ICA standards, the software development project of ICA (*ICA Atom*) shows a particular perspective over relationships between entities described (Appendix 8). On the top of entities there is a taxonomy of terms that provides controlled vocabulary for other entities. The entities are *archival institutions*, *actors* and *archival material*. One can notice the absence of ISDF, but it seems to be only a mistake as it is taken into consideration as part of ISAAR records.

SPANISH MODEL[14]

Based on ICA standards, a very interesting and complex approach has been developed by Spanish archivists and published in 2011. In the document archival entities and their relationships are presented: Agents, Functions, Mandates, Records, Concepts/Subjects and Places. For Agents, Functions and Records there are defined inner hierarchical relationships.

HURLEY'S COMMON PRACTICE RULES (HCPR)[15]

Compiled and released in 2009 by Chris Hurley, HCPR presents an Australian model for archival description. It has three entities (or rather entity types): Deeds, Documents and Doers. *Deeds* "refers to what are variously termed functions, activities, business activities, actions, mandates, authorisations, business, recordkeeping, relationships, and acts"; *Documents* "refers to what are variously termed fonds and sous-fonds, record groups, series and sub-series, sequences and super- or sub-sequences, items, files, documents, documentary objects, accessions, consignments, transfers, sets"; *Doers* "refers to what are variously termed organisations, agencies, persons, families, corporations, agents, actors, institutions with archival holdings, libraries, museums, collections, galleries, custodians".

LOCAH Project[16]

One last model that I present here is the one generated within the LOCAH Project. This project aimed to connect two description catalogs (COPAC and Archives Hub), in order to make available their information as *linked data*[17]. The model it results, based on Encoded Archival Description, is presented in Appendix 10. The model starts from the entity *Archival resources*, that is directly connected to *repository* (type of Agent), *Finding Aid*, *Biographical History*, *Agent*, *Concept*, *Books* or *Object*. Although very specific, its approach is interesting since it identifies many various entities that might be relevant for an archival resource.

---

13 https://www.ica-atom.org/doc/Entity_types.

14 *Modelo Conceptual de Descripción Archivística y Requisitos de Datos Básicos de las Descripciones de Documentos de Archivo, Agentes y Funciones, Parte 1: Tipos de Entidad, Borrador final de la CNEDA (15-12-2008) at* http://www.mcu.es/archivos/docs/NEDA_MCDA_P1_P2_20110609.pdf.

15 *http://www.descriptionguy.com/images/WEBSITE/hcpr-hurleys-common-practice-rules-2009.pdf.*

16 http://archiveshub.ac.uk/locah/2011/02/16/two-changes-to-the-model-and-some-definitions/.

17 http://archiveshub.ac.uk/locah/about/.

Some remarks over these data models

This short summary of data models related to records and archives did not intent at all to be exhaustive. Also, a rigorous reader might find other issues about the consistency of various models: one is entity-relationship model, other is object-oriented; one is designed for general metadata, other only for preservation purposes and so on. It was not our intention to sharply categorize models, but to have a broader perspective of how (differently) professionals see the abstraction of their work object.

Despite these limitations, some remarks can be done.

Firstly, in almost all models three entities are commonly present: **records**, or archival materials, **agents** or actors or creators; **functions** (business, mandates and activities). It seems there is a general agreement on the importance of these three entities when dealing with records and archives. They seem to be, at date, the pillars of archival data models.

Secondly, models are designed either for current records or historical archives, but many common elements can be identified. Although for a records management model it is unlikely to discuss the topic of an *archival fonds*, the historical archives approach might take into consideration the entity of a *record*, as an item. This marks a profound shift in our profession today. Digital environment tends to change one classical difference in managing records or archives: if, in the past, records management dealt with individual entities, while archives management dealt with large groups of records (collective entities), nowadays the proper electronic records management will transfer also its control tools to the historical archives. That means the the nature of a record (as a discrete element) is as important for historical archivists as the nature of an archival fonds. Moreover, the classification scheme for records is of the utmost relevance for arrangement of historical archives: it gives the full context of creation, it helps in retrieval and might explain why the same record (read *content*) might be identified in many classes or folders.

In the third line, looking on the example above one can notice a lot of entities that, in different models, are considered to be significant. Are they generally relevant or they are only for parochial use only? Is indeed the entity "Places" a core entity type for an archival data model? Or it is rather an access point for information within archival fonds?

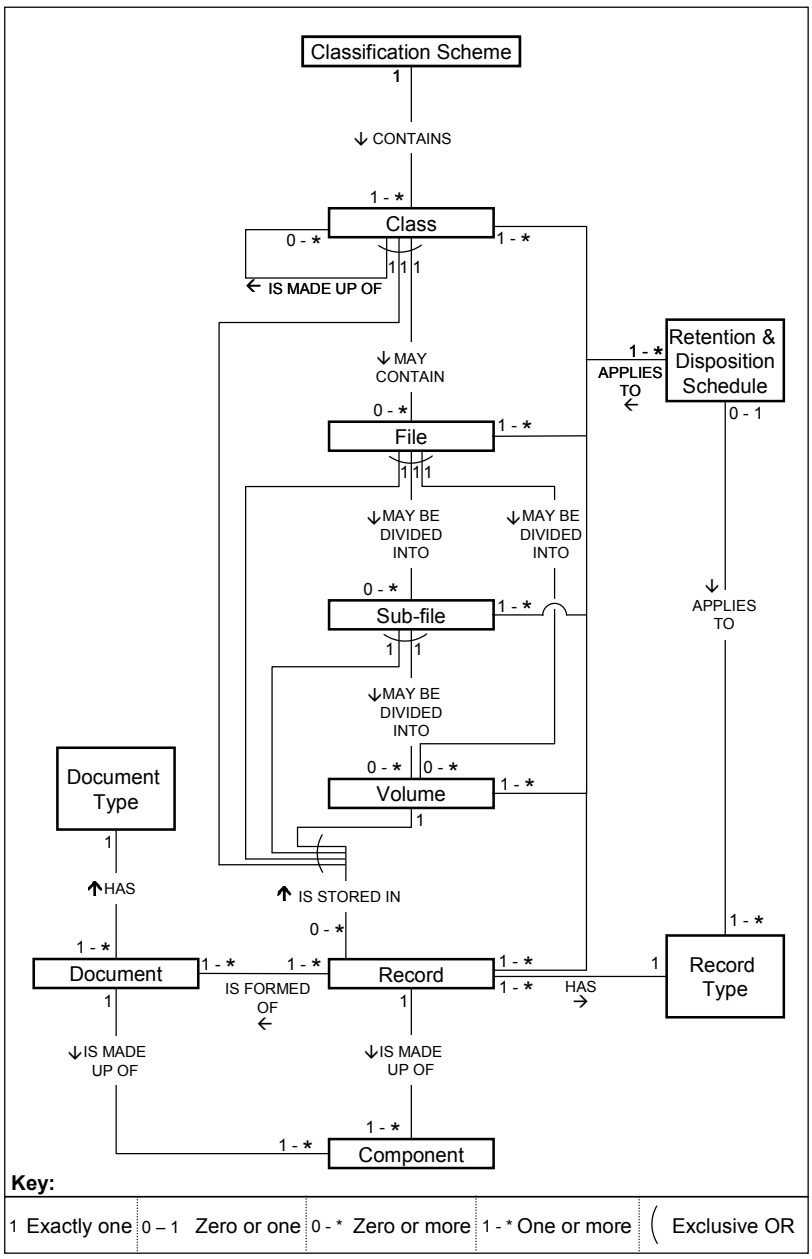
As a last remark I would like to argue what I like to call "challenging the provenance". Since the dawn of modern archival science (or even before), Principle of Provenance and that of Original Order was the backbone of our profession. Arranging records into aggregations (*fonds*) based on their creator and respecting the order the creator gave to those records was the best and most "objective" option in comparison with a subjective arrangement of an archivist. But today, when digital environment makes everything seem so easy, Original Order and Provenance might rest more and more only as a *possible* arrangement. With the click of a button "content" might be arranged on dates or topics or places — it is all virtual. Access points I mentioned before may claim their place as a reasonable criteria for arrangement. This is especially true when the records accessioned to historical archives might be in a system compatible with MoReq2010, with multiple classifications and aggregations. Which is here the original order then? Is it not more like a game, a puzzle of metadata, where each user might choose her/his arrangement?

In this respect, I would like to add another argument. All the "classical" archival material description standards emphasize hierarchy. But, going into an archival repository, there is no hierarchy; in fact, there are horizontal rows of records and/or files. It is only us, humans, that assign to those records/files different metadata in order to aggregate them into complex hierarchies, into series, sub-fonds or fonds. If a file is mis-arranged and afterwards it is put in its proper place, it, as an object, is unchanged; but suddenly, it will become a part of another "fonds", it will get new attributes and new meanings. All is metadata, the only entity is record itself. It is at this level where all the necessary data for context, content and structure should be defined. Upper levels are only groupings of similar metadata and we called them aggregation of records.

Another challenge of provenance might be the representations. Traditionally, multiple copies of records are eliminated, since one copy had the necessary historical information. That unique copy is part of a fonds. But the Archives, as institutions, digitize today a lot; or generate working copies of a digital born record. Are these copies part of the original fonds or not? If we consider them as a part of the same fonds, one may notice the creator of the copy is not the creator of the fonds and this is breaking the Principle of Provenance. If it is considered as new product, copies must be treated as new digital records of the Archives. Are these new virtual collections of "representations" entities of an archival data model? Or are they part of a larger batch of entities the Archives are dealing with?

I have not proper answers for all the questions above. But it is relevant that many professional communities answer the same questions differently. Maybe it is time for the profession to ponder to an overarching data model, to reinterpret traditional principles and give them a more modern meaning, consistent with dramatic changes in the creation of records that occurred in the last decades.

Appendix 1: MoReq2 entity model



Appendix 2: MoReq2010 and some new approaches

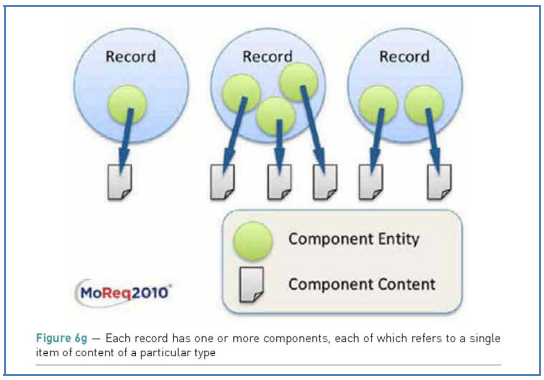


Figure 6g — Each record has one or more components, each of which refers to a single item of content of a particular type

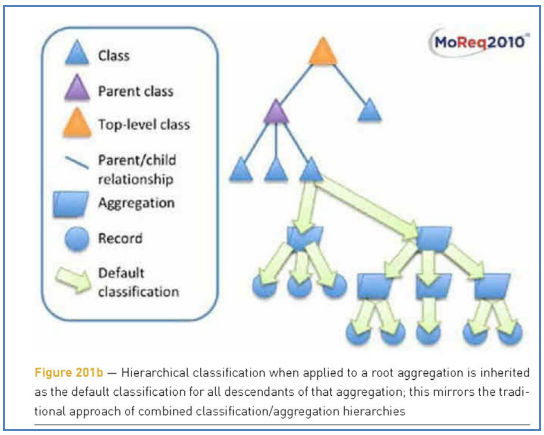


Figure 201b — Hierarchical classification when applied to a root aggregation is inherited as the default classification for all descendants of that aggregation; this mirrors the traditional approach of combined classification/aggregation hierarchies

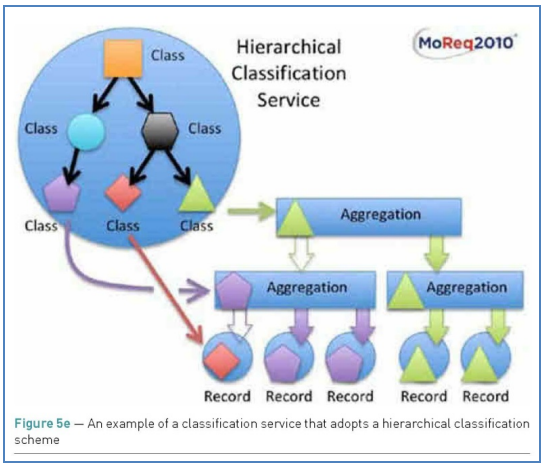


Figure 5e — An example of a classification service that adopts a hierarchical classification scheme

Appendix 3: CIDOC Conceptual Reference Model for public sector information

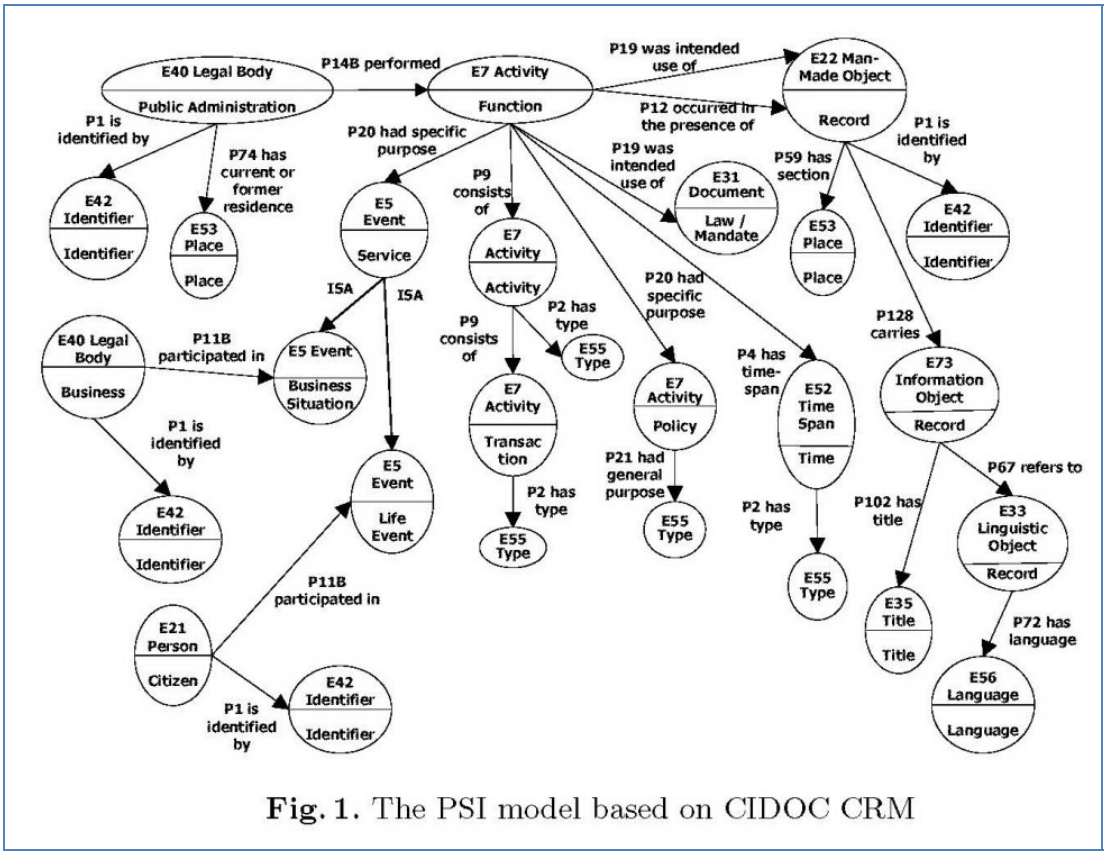


Fig. 1. The PSI model based on CIDOC CRM

Appendix 4: PREMIS data model

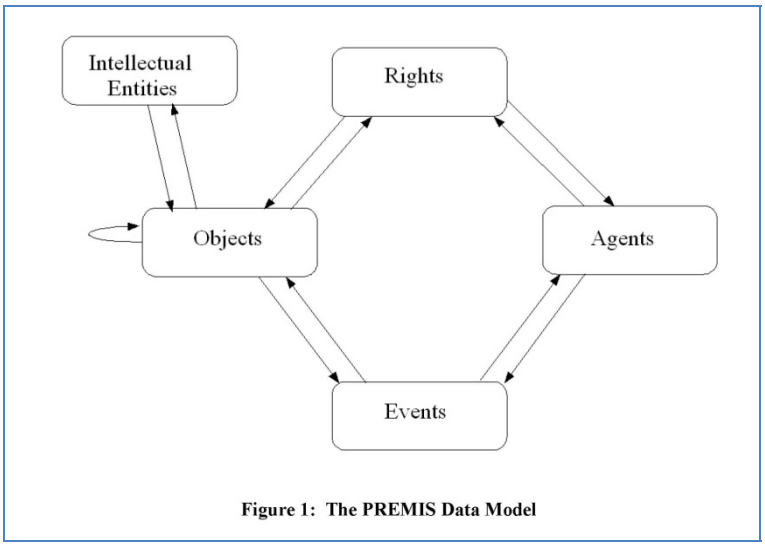
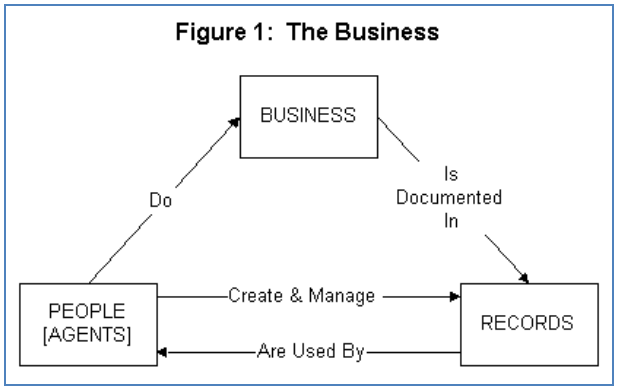
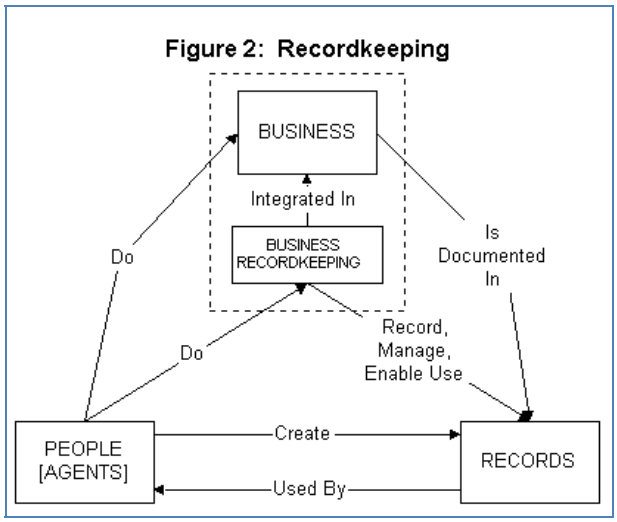
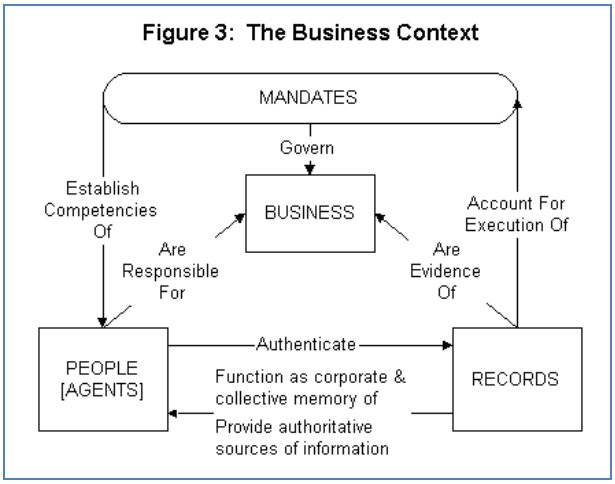


Figure 1: The PREMIS Data Model

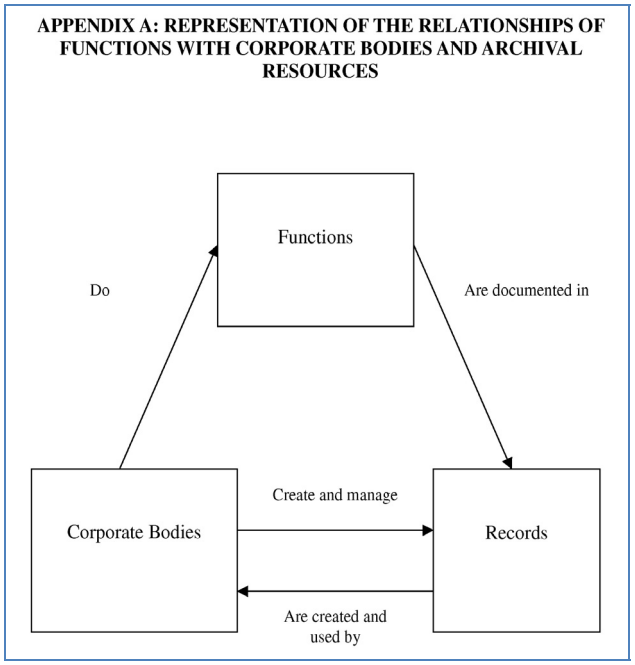
Appendix 5: Australian recordkeeping metadata model



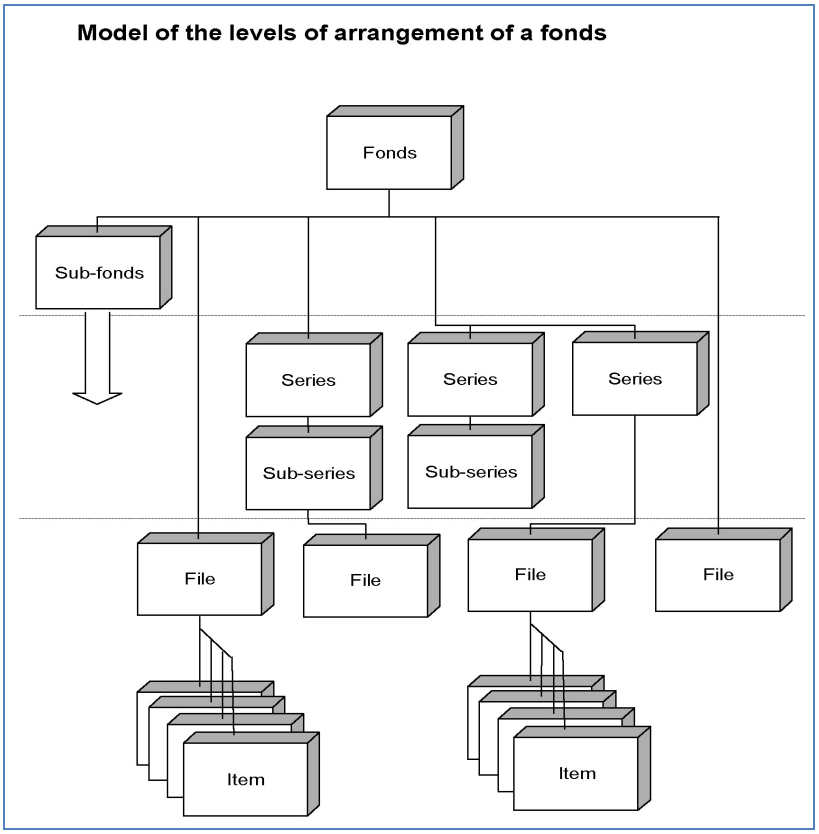




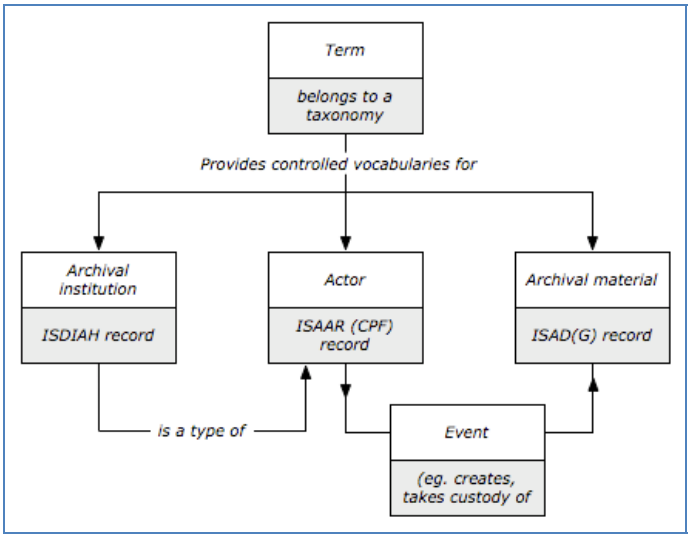
Appendix 6: ISDF diagram



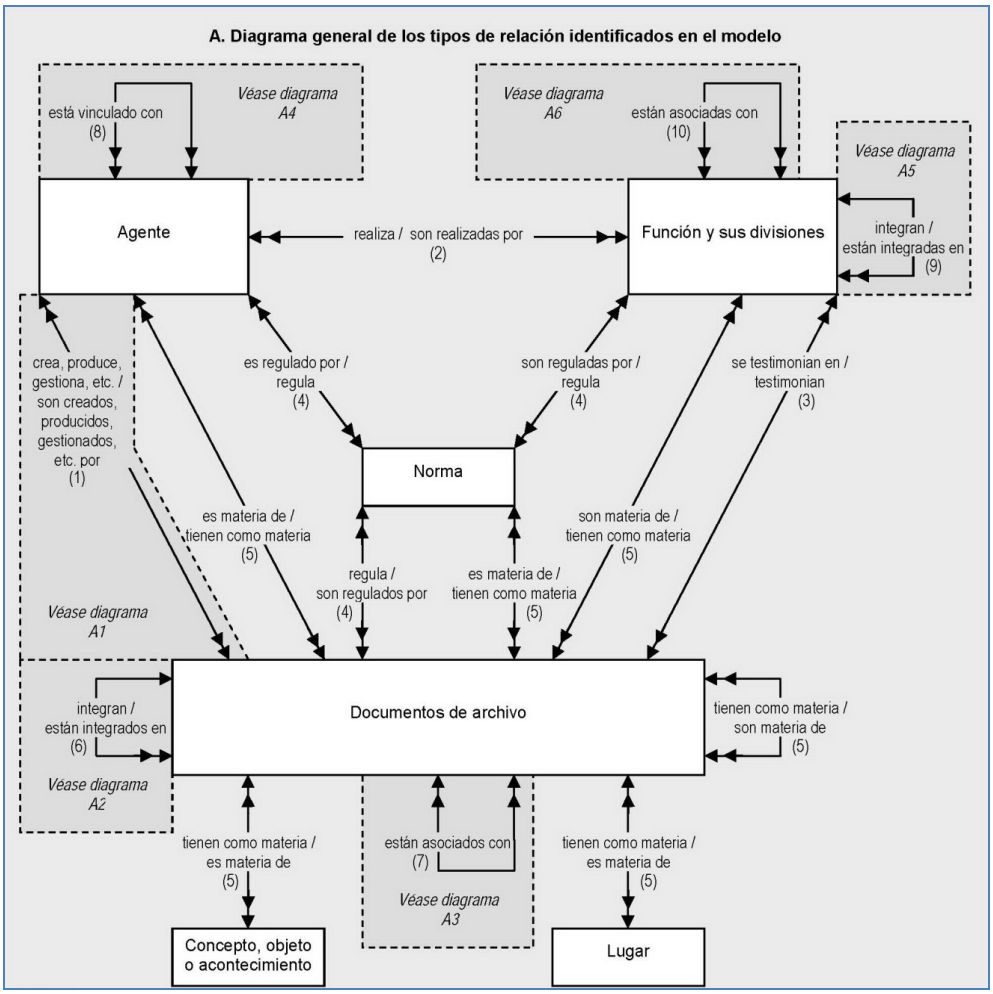
Appendix 7: ISAD(G) hierarchy



Appendix 8: ICA Atom entities



Appendix 9: Spanish archival model



Appendix 10: Entity model from LOCAH Project