Records Management Support in the Interoperability Framework for the Portuguese Public Administration

Catarina Viegas[1], André Vasconcelos[1,2], José Borbinha[1] and Zaida Chora[2]
[1]*INESC-ID, Instituto Superior Técnico, Avenida Rovisco Pais 1, Lisbon, Portugal*
[2]*Administrative Modernization Agency, Rua Abranches Ferrão 10, 3, Lisbon, Portugal*

Keywords: Information Management, Public Administration, Interoperability, Records Management Metadata, Canonical Data Model.

Abstract: The Portuguese public administration has a core technological infrastructure for interoperability, which assures reliable core transactions, but takes all information objects as equals, leaving any necessary specialization to the applications. However, public administrations are highly regulated environments, which implies business processes involving entities of that domain are subject to strong requirements for information management. Records management in special is a specific concern, meaning metadata for that purpose must be produced along the production of the regular business information objects. In that sense, when two or more entities of a domain of this kind engage in transactions, it is helpful for all those involved if also metadata created for that purpose can be shared, which requires it to be commonly understood. In Portugal, national guidelines have been developed to support that goal, remaining now the challenge of their implementation. This is a classic problem of interoperability in distributed information systems, which has particular challenges when scoped in the domain of a large public administration, involving thousands of local systems. This paper describes the results of a research project intended to provide a proof of concept for that for the case of the Portuguese public administration, which resulted in a case of application of the Canonical Data Model method. The metadata schema produced is assessed using the Bruce-Hillman metadata quality framework, which made possible to conclude by its effectiveness, along with suggestions for future improvements.

1 INTRODUCTION

Organizations within the same public administration exchange information frequently among them. This information, which used to be mainly in the form of physical paper-based documents, tend to be now business objects when the transactions are supported by digital information systems. However, these business objects need to be kept as records in the sending and recipient organizations.

Records are evidences of processes, making the management of records within an organization extremely important. The Portuguese Public Administration (PPA) is made of multiple organizations, each one expected to manage its records according to its specific regulations and requirements, while obeying to a same general legal framework. For that purpose, all entities are expected to have defined specific records management systems (RMS), conceived to capture, store and manage records (Barbedo and Corujo, 2012).

The promotion of measures by the Portuguese government to dematerialize business processes, led to the development of an interoperability project which could ensure the sharing of information through the RMS of public organizations.

The existence of an interoperability infrastructure for the integration of information systems for the PPA, and common requirements for records management previously defined, motivates a solution for this project, based on existing infrastructures.

This paper presents the results of the development and validation of a data model for records metadata. This document follows by presenting a description of the current interoperability measures in the PPA, and an overview of the techniques for ensuring systems interoperability. These techniques are the basis for the development of the solution, which is presented in section 3. The results of the experiments performed on the solution proposed are analyzed in section 4 and the conclusions and future work are presented in section 5.

Viegas, C., Vasconcelos, A., Borbinha, J. and Chora, Z.
Records Management Support in the Interoperability Framework for the Portuguese Public Administration.
DOI: 10.5220/0007673300840094
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 84-94
ISBN: 978-989-758-372-8

2 BACKGROUND

This section presents the most relevant research developed regarding the management of records, interoperability and their current role in the PPA. In this section, are also presented the different integration approaches considered for this project.

2.1 Records Management and Records Management Systems

According to the ISO 15489-1:2016 (ISO 15489-1:2016, 2016), **records** are information that is created, received and maintained as evidence of an organization's business process and as an asset in pursuit of legal obligation or in the transaction of business.

The same standard defines **records management** as the *"field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records"* (ISO 15489-1:2016, 2016), and thus accordingly it defines **Record Management Systems** as information systems that capture, manage and provide access to records. According to the same standard, all records should be defined by metadata elements and every system should have one or more metadata schemas which state how to define a record.

Related work has been described in (Maguire, 2005), which depicts the implementation of a records management system in the Estates Department of the British Library, describing the decisions made throughout its implementation process, including the definition of an adequate metadata schema.

2.2 Interoperability in Public Administrations

The Decision no. 922/2009/EC (European Parliament and of the Council, 2009) defines interoperability as the capability of two or more diverse public administration (PA) organizations to interact by sharing information and knowledge through the exchange of data between their information systems. The European Interoperability Framework for European Public Services describes four levels of interoperability (European Commission, 2010): legal, organizational, semantic and technical. This research focus mainly on achieving technical and semantic interoperability across the PPA, promoting, consequently, the other two levels.

Semantic interoperability is the capability of two or more information systems to exchange information, while guaranteeing that the information's original meaning is maintained after the exchange, in the recipient system. The exchange of data across different information systems can face multiple barriers, such as the lack of a commonly agreed metadata schema, or divergences in interpretation of the data exchanged (European Commission, 2017). Therefore, the establishment of a common reference to be used by every organization is crucial to achieve interoperability in the domain of PA. The definition of a metadata schema that is used by every organization within the PA will facilitate the correct sharing of metadata records every time two RMS engage in a transaction. A metadata record is shared when it is produced in a RMS and sent to another RMS, and reused by the receiver to create a local record. MIP[1] is the current Portuguese metadata schema produced to be applied by PA entities when managing their records.

2.3 MIP - Metadata for Interoperability

DGLAB[2], the entity that has the role of national archive in Portugal, defined MIP to support metadata interoperability for records management, with the goal of defining a common schema to be used by public agencies to characterize their records.

MIP is a metadata schema, comprising 17 metadata elements, defined to ensure semantic interoperability within the PPA (Barbedo and Corujo, 2012). By defining a common schema to be applied by all different PA entities to their records, the goal was to ensure that the data exchanged was equally interpreted by every RMS of the PPA. This way, local records, copies of the records in the original RMS, could be automatically created in the recipient RMS, aided by the data received. To identify the metadata elements important to be in the schema, requirements from records management international standards were considered. These standards state the metadata elements each record should contain to guarantee the record's authenticity and reliability (Barbedo and Corujo, 2012).

Even though the development of MIP was promoted by the Portuguese government, it is not legally mandatory for PPA organizations to use it in their records. This generates a semantic problem across organizations. The use of the same data schema to characterize records guarantees that records are rightfully recognized, captured, stored and managed by any system that supports the schema, achieving RMS inter-

[1]"Meta-informação para Interoperabilidade" in Portuguese.

[2]"Direção Geral do Livro, Arquivos e Bibliotecas" in Portuguese, which stands for General Directorate for Book, Archives and Libraries.

operability.

2.4 MEF/LC - Functions and Processes in the Portuguese Public Administration

MEF/LC is the result of the national project ASIA[3] (Lourenço and Penteado, 2015) and consists in the merge of MEF[4] and LC[5]. MEF is a classification scheme that constitutes a conceptual representation of the functions performed by public sector organizations, providing two levels of classification. The first level represents the state functions and the second the subfunctions in which the level 1 instances can be divided (for example, "Strategic planning and management"[6] is a state function of level 1, with code 150, and "Policy definition and evaluation"[7] is a subfunction with code 150.10) (General Directorate for Book, Archives and Libraries (DGLAB), 2013).

LC (General Directorate for Book, Archives and Libraries (DGLAB), 2014) is a catalogue of the business processes executed by the PPA. MEF/LC collects the information provided by these two classification models, establishing a 4-level classification scheme to be used by the organizations as a referential in the development of their own functional business classification schemes (Lourenço et al., 2012). MEF/LC is summarized by a table of codes that define the functions, subfunctions and business processes executed by public agencies.

The main problem with MEF/LC is the lack of mandatory legislation able to establish the use of this classification model by public organizations, in similarity with MIP. Since it is not of mandatory use, only a small number of organizations of the PPA use this classification model in their records. As a result, organizations can choose to apply this or other classification model, generating a discrepancy in the way records are classified.

Multiple services offered by the PPA require a collaboration between different public entities. This collaborative approach is often achieved through the exchange of documents among organizations, reason why interoperability has always been a concern in the PPA. Interoperability measures such as MIP or MEF/LC were developed considering this collaborative feature of the PPA. However, a closer analysis of these measures allowed to conclude that they may not be enough for the scope of this research. This work will provide the PPA with an interoperability solution that considers the measures defined, whilst being capable of mitigating the flaws they may possess.

[3]"Avaliação Suprainstitucional da Informação Arquivística" in Portuguese.

[4]"Macroestrutura Funcional" in Portuguese

[5]"Lista Consolidada" in Portuguese.

[6]"Planeamento e Gestão Estratégica" (in the original (General Directorate for Book, Archives and Libraries (DGLAB), 2013))

[7]"Definição e Avaliação de Políticas" (in the original (General Directorate for Book, Archives and Libraries (DGLAB), 2013)).

2.5 Integration of Information Systems

In this subsection we will introduce the fundamental state of the art of architectures for integration of information systems, in relation to the the integration platform currently used by the PPA.

2.5.1 SOA and ESB Architecture

Nowadays, where integration is concerned, businesses opt for approaches like a Service-Oriented Architecture (SOA) and an Enterprise Service Bus (ESB). SOA provides the capability of designing the business as a collection of application, where each is responsible for one task within a business context. ESB are integration platforms that allow the coordination of the interaction between different applications from different sources (Chappell, 2004), by routing messages from one application to another.

Often used together, these approaches may encounter limitations, specially regarding data integration. Data integration is the process of combining different data from various sources to generate a unified view of all the data intended. Data integration can become very complex when using a SOA-ESB approach, specially in large SOA projects, since multiple systems exchange data with one another but can have different data definitions. To mitigate this problem, ESB offers message transformation, the process of converting the data format of a message to another, through the definition of mappings that correlate the different data schemas with each other, which can be applied in Point-to-Point Integration pattern or a Canonical Data Model approach.

2.5.2 Point-to-Point Integration

A Point-to-Point integration technique requires, for each service, the manual creation of a message translator for every application it interoperates with, establishing a translation per interaction. Therefore, each different data schema is translated as many times as there are different data schemas within a SOA. Any changes in any of the data schemas implies changing its translation in every system it communicates with. Figure 1 (left) depicts the number of message

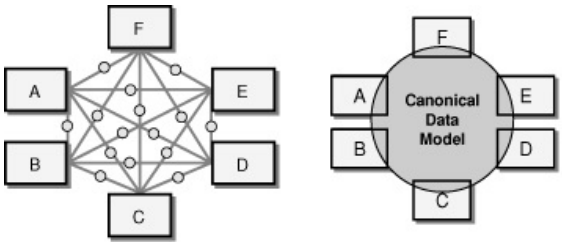

Figure 1: Point-to-Point Integration (left) and Canonical Data Model (right) (Hohpe et al., 2004).

transformations that would be required in a process with six different applications and six different data schemas.

2.5.3 Canonical Data Model Integration

The Canonical Data Model (CDM) methodology consists of developing a data model known by every system within the same ESB process. Each system is responsible for defining their own message translator, from the CDM to its own data format, to interoperate with another application of the ESB. This way, every system only needs to develop one message translator, instead of developing a message translator per different data format within the ESB. Figure 1 (right) illustrates how a CDM supports the integration of an application within the ESB. This approach ensures that applications are able to interoperate with one another as long as they are able to translate the CDM into their own data format.

2.5.4 Comparative Analysis

Regarding the number of translation steps, Point-to-Point Integration approach requires only one translation step in an interaction between two different applications, whilst the Canonical Data Model Integration approach requires a double translation (Hohpe et al., 2004): from the source application's data format to the CDM and other from the CDM to the target application's data format. Considering that each translation adds latency to the message flow inside the ESB (Hohpe et al., 2004), the need to introduce an extra translation step can decrease performance results in a Canonical Data Model approach.

Point-to-Point integration also requires a different message translator for every application, which increases the complexity of this process every time a new application is added, or every time there is a need to change the translation of a data format. The use of a CDM reduces the complexity of message transformation, and still guarantees the flexibility and heterogeneity that characterize a SOA (Dave Hollander, 2011). With a CDM approach, a new level of indirection is added among applications' individual data formats, making easier future changes in the data format (being only necessary to update that application's translator) and the integration of new applications (since there is only the need to implement a translator for the application and the common model and not a translator for every application).

Regarding scalability, the two approaches present limitations when a change occurs. In a Point-to-Point integration, if there is a change in any of the data formats, all translations need change. In a CDM approach, if a change in the common data model is required, every application needs to update their data transform. On the other hand, the CDM can reduce the complexity of this process by guaranteeing that, when developing the model, the process is done with the maximum level of abstraction possible, considering every application's data format, while still ensuring a response to the business needs.

The use of a Point-to-Point integration is difficult in large SOA projects, since it requires a large number of message translators. Considering the diversity and size of the PPA, the definition of a CDM was the approach chosen. Even if sacrificing the performance, by adding the second translation step required by the CDM, the reduction of the complexity of the process pays off in the end.

2.5.5 iAP - Interoperability Framework for the Portuguese Public Administration

The Administrative Modernization Agency (AMA), is the Portuguese agency responsible for promoting modernization within the PPA. In that scope, it developed iAP[8], the interoperability framework for the PPA. The main objectives for this development were to 1) simplify the communication between organizations and business partners by streamlining business processes and developing services, and 2) facilitate and minimize the costs and effort of developing new business processes (Administrative Modernization Agency (AMA), 2011). Although the platform has four main components, only the functionalities provided by the Integration Platform, a platform developed as a state-wide SOA, that provides a catalog of services published and consumed by entities of the PPA, will be explored. The invocation of these services is mediated through the use of an ESB, provided by iAP.

[8]"Interoperabilidade da Administração Pública" in Portuguese.

3 SHARING RECORDS METADATA IN THE PORTUGUESE PUBLIC ADMINISTRATION

Interoperability can only be achieved through a mutual agreement on all basis: technical, semantic, organizational and legal. The solution developed focus on technical and semantic interoperability, in the development of a consensual data model to be applied as the service interface for the new service in iAP, which allows the exchange of metadata. A service interface defines which data the service needs, describing the message format to be used for data exchange among systems. A service interface can be defined by a WSDL or an XSD, reason why the CDM developed was implemented as the XSD of the iAP service.

3.1 Solution Overview

iAP offers a set of services, all of them defined by a specific CDM, ensuring that all the information required is provided by the organizations who invoke these services. The solution developed is a CDM for a new service in iAP, which will allow the sharing metadata among systems.

The development of this CDM was based on MIP (see subsection 2.3). Although MIP has issues, as stated and shown next in subsection 4.1, the information it provides must be preserved and included in the CDM proposed. Using MIP's element definitions, each element of the CDM has obligation and repeatability attributes that state if the element must be present and if it can appear more than once in the SOAP message generated when the service is invoked.

This section will be divided into subsections, each one representing a different type of change MIP elements suffered when represented in the CDM proposed.

3.1.1 Equivalent Element Definition

This subsection refers to the elements that maintained the definitions proposed by MIP, when represented in the CDM.

Elements *Aggregation* (*Agregacao*), *Subject* (*Assunto*), *Coverage* (*Cobertura*) and *DocumentType*'s (*TipoDocumental*) structure and meaning remain identical to MIP, the only difference being the way a document type is represented. In the CDM, rather than following MIP's definition of the element, by allowing the designation of any value, a numeric code was assign to each document type considered by this research.

3.1.2 Addition of New Subelements

Although MIP specifies the elements necessary for a correct description of the record, it was detected, throughout this work, a need to add subelements to already defined MIP elements, to complete the information provided.

Element	Subelement added	Mandatory?	Repeatable?
Description	ScopeandContent	No	Yes
Title	FormalTitle	Yes	No

Figure 2: Elements with new subelements.

As shown in Figure 2, *Description* (*Descricao*) and *Title* (*Titulo*) are two MIP elements to which new subelements were added. *ScopeAndContent* (*AmbitoeConteudo*) and *FormalTitle* (*TituloFormal*) were added to these elements to complete the information already provided.

3.1.3 Deprecated subelements

Considering the research context and the circumstances in which MIP was developed, some of the subelements presented in it are no longer valid or necessary for the CDM. The reasoning behind this decision is either their insignificance for the research (MIP subelement *Support*), or the redundant information provided by them (remaining subelements referred in Figure 3).

Element	Subelements removed	Mandatory?	Repeatable?
Format	Support	Yes	Yes
RecordDates	Start date; Availability date; End date	Yes	No
Accessibility	Digital signature authentication; Access list	Yes	No

Figure 3: Elements that lost subelements.

Represented in Figure 3[9], elements *RecordDates* (*DatasRecurso*), *Format* (*Formato*) and *Accessibility* (*Acessibilidade*) have all lost one or more of their associated subelements represented in MIP.

3.1.4 Redefinitions of Subelements

In the development phase of the CDM, it was defined that the information transmitted by MIP ele-

[9]The English translation of the names of the MIP subelements presented in the table are of the responsibility of the authors. The original names, in Portuguese, can be consulted in (Barbedo and Corujo, 2012).

ments would be equally transmitted by CDM elements. However, changes to the elements were required to ensure that the original meaning of this information was maintained when captured by another RMS.

The information provided by these elements is the same, but the way it is structured is different, in order to respond to the needs of automation of metadata capture and record creation processes, promoted by the proposed solution.

For identifying organizations, elements *Identifier* (*Identificador*), *Producer* (*Produtor*) and *Receiver* (*Destinatario*) all apply the same structure to their subelements. *XOrganizationType* (*TipoOrganismoX*) and *XOrganizationID* (*IDOrganismoX*) are two types of subelements that were introduced with the goal of providing a normalization to the identification of organizations. X represents the role of the organization.

The way organizations are identified in the CDM changed when compared to MIP. The proposed CDM considers the need for an automation of the processes of capturing metadata and registering new local records in the RMS. This automation can only be guaranteed if the data received is normalized. Thus, it was defined that organizations would be identified using only one common method, a SIOE[10] code. SIOE assigns to each organization an unique code. The type of organization is stated in element *XOrganizationType*, with values raging from 1 to 3 (1 is a SIOE registered organization, 2 is a non-SIOE organization but iAP-user and 3 is for the remainder that do not fall in none of the other categories. The unique number that characterizes the organization is hold by element *XOrganizationID*.

3.1.5 New Metadata Elements

When developing the CDM, new metadata elements were also introduced. Figure 4 displays all of the elements that are not represented in MIP and were added to the CDM. The obligation and repeatability of the elements is represented in the table and both features will affect the way they are implemented in the XSD of the service, as shown in subsection 3.2.

Element	Subelements?	Mandatory?	Repeatable?
KnownReference	Yes	No	No
Priority	No	No	No
SpecificMetadata	Yes	No	Yes
OtherClassificationCode	Yes	No	Yes

Figure 4: New elements introduced in the CDM.

The element *Priority* (*Prioridade*) was added to provide the organization with the information regarding the handling priority of the data received. *KnownReferece* (*VossaReferencia*) is used only when the metadata is sent as a response to a request. An information object B is considered a response when its contents answers the contents of information object A. This element is exclusively meant to indicate which related object triggered the response object being sent. By using this element, the recipient RMS would recognize B as being a response to A, independently of how many objects were related to object B, identified in element *Relationship*, a MIP element who transitioned to CDM, whose functionality is stating which records are related and the type of relationship they established with one another.

SpecificMetadata (*MetadadoEspecifico*), is a new subelement which provides organizations with a way of adding extra information in the metadata. *OtherClassificationCode* (*OutroCodigoClassificacao*) was introduced to provide organizations that do not use MEF/LC as a classification model, the capability of identifying the classification model used. The addition of this element was a necessity considering the CDM was designed to promote the use of MEF/LC by the organizations of the PPA.

3.2 XSD Implementation

For implementation, the new data model was applied to the XML Schema Definition (XSD) of the iAP service.

To represent the repeatability and obligation of the elements, the `minOccurs` and `maxOccurs` indicators are used to specifies the minimum and maximum times an element has to appear in the schema. To represent a repeatable element (i.e. an element that can appear more than once in the SOAP message generated), `maxOccurs`' value must be `unbounded`. To represent a mandatory element, indicator `minOccurs` must take the value of 1 if the element is mandatory, and the value of 0 if the element is optional. Listing 1 depicts an excerpt of the XSD file where element *OtherClassificationCode* is characterized as being optional (`minOccurs = 0`) and repeatable (`maxOccurs = unbounded`), as shown in Figure 4.

In the XSD, elements are portrait as `complexTypes` if they have associated subelements, or `simpleTypes` if not. Subelements who have subelements, are represented by `complexTypes` as well, generating an hierarchical architecture that allows systems to comprehend which elements cannot exist without their parent elements.

Elements such as *Relationship* or *Identifier* have subelements whose range of accepted values is lim-

[10]"Sistema de Informação da Organização do Estado" in Portuguese.

ited. *RelationshipType* (*TipoRelacao*) is a subelement of *Relationship* that represents the type of relationship established between two records. All types of relationships considered are represented by a numeric code, ranging from 1 to 12. To represent this limited set of values, the `restriction` attribute is used. As shown in Listing 2, element *RelationshipType* can only accept integer values ranging from 1 to 12, implemented by the use of `minInclusive` and `maxInclusive` indicators.

Listing 1: Excerpt of the service's XSD file for element occurrence.

```
<xs:element name="OtherClassificationCode"
    type="OtherClassificationCode"
    minOccurs="0" maxOccurs="unbounded"/>
```

Listing 2: Excerpt of the service's XSD file for value restriction.

```
<xs:element name="RelationshipType"
    minOccurs="1">
        <xs:simpleType>
            <xs:restriction base="xs:int">
                <xs:minInclusive
                    value="1"/>
                <xs:maxInclusive
                    value="12"/>
            </xs:restriction>
        </xs:simpleType>
</xs:element>
```

4 RESULTS AND DISCUSSION

To evaluate the proposed CDM, two methods were employed: 1) a comparison between MIP and the proposed CDM, to determine and analyze the importance of the CDM and the improvements presented by the model, regarding MIP; and 2) an assessment of the qualities possessed by the CDM developed, according to the Bruce-Hillman Framework (Bruce and Hillmann, 2004).

4.1 Comparing MIP and CDM

The comparison between MIP and the CDM developed consisted on the application of the two data models as metadata schemas of iAP's service, in the same scenario. By applying MIP as the metadata schema, two types of problems were identified: 1) Lack of rigid norms of application (Problem A) and 2) Structural problems (Problem B). The CDM presents solutions for these faults, as it will be described in this section.

Problem A is summarized by the lack of controlled vocabularies and limits to the range of values of the elements in MIP. Controlled vocabularies are a set of accepted values that metadata elements can hold (Online Computer Library Center, 2013). MIP documentation provides examples of the values elements can possess. However, this is not enough to guarantee interoperability, because it is not certain that every organization will use and understand the values equally. With MIP, every organization can apply any value they deem fit whilst, to ensure semantic interoperability among different systems, every organization must apply the same set of values defined. This lack of common values also hinders the automation process. The CDM proposed introduces a set of controlled vocabularies to provide organizations with a limit set of valid values for the elements proposed.

Problem B arises from the structural problems of MIP. These structural problems were found in three crucial MIP elements. MIP element "Identificador de recurso" (Record Identifier) is responsible for identifying uniquely a record. The problem with this element is the way it is structured in MIP, which does not ensure the uniqueness of the identifier of a record within an interoperability process. MIP element "Relação" (*Relationship*) also has structural problems in subelement "Tipo de Relação" (Relationship type), which contains multiple subelements, one for each type of relationship the record may established with another. This structure is not efficient since most of the subelements would be left blank.MIP element "Código de classificação"'s (Classification code) structure constitutes a problem, with users not being able to identify the classification model used to classify the record and, consequently, hindering the interpretation of the record's class by the receiver system.

For a better understanding of the origin of these problems and how the CDM provides solutions for them, an example will be presented.

As illustrated by Figure 5, suppose that Entity A wants to inform Entity B. Upon receiving the document file and metadata (represented in Figure 5 as an unique entity for a simple understanding), a local record is created in the RMS of Entity B, from the information received. After analyzing the received information, Entity B develops a response document, registers the related information in its RMS and sends the response to Entity A. Without the solution proposed, entities would send the document via e-mail or letter, which would lead to unnecessary costs and time spent, and could potentially lead to an ineffective management of records, since there is no guarantee that a new record would be locally created in the

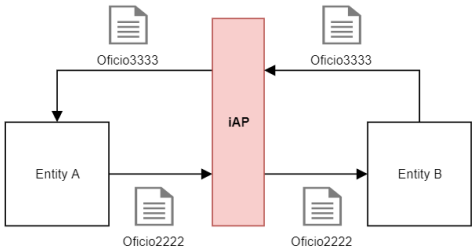

Figure 5: Example of flow of documents through iAP.

RMS of Entity A upon receiving the document.

As stated, the use of MIP AS-IS as the canonical data model of the service would lead to multiple faults. The way records are identified using MIP would generate a problem if used to identify records in the context of this work. MIP offers only one element to hold the value identifier of the record. This element in MIP does not have any special structure to ensure that the identifier of the record, which must be unique, remains unique when shared with multiple organizations within the same interoperability process. To solve this issue, CDM introduces a new structure for the identifier element, with three subelements. Subelements *GeneratorOrganizationType* and *GeneratorOrganizationID* provide information about the organization that identifies the record, while *DocumentID* holds the identification code provided by said organization. This way, even if two entities have different records with the same identifier, as each organization is identified with an unique code, the set of the three subelements maintains the record's unique identifier.

To identify the classification model used, MIP uses one single element, which can hold any value the organization deems appropriate, without restrictions, since MEF/LC is not of mandatory use. However, the key for success resides on the receiver organizations identifying correctly the class of the record. Without knowing which classification model is used, that is not possible. Thus, when developing the CDM, an extra element (*OtherClassificationCode*) was introduced, to provide non MEF/LC compliant organizations of informing the receiver system of which classification model is being used. However, the use of MEF/LC is promoted by the CDM, through the restructuring of element *ClassificationCode*.

The use of MIP in this context generates a problem in the way organizations are identified in the metadata. In its documentation, MIP provides examples of values for identifying organizations. The problem is that these examples are considered as a suggestion only, and not as a norm, leaving to organizations the responsibility of using any method they deem fit for identifying organizations, not guaranteeing a semantic agreement across all organizations. Thus, in the CDM, it was defined that every entity would be identified using SIOE or, if not present in SIOE, other code that is on a database controlled by AMA, as stated in subsubsection 3.1.4.

To identify the relationships established with the record, MIP provides element "Tipo de Relacao" (Relationship type), which contains 11 subelements, each one responsible for identifying a different type of relationship. As stated, this is not efficient. The CDM provides a new structure for identify the relationship type, assigning to each type a numeric code, to populate subelement *RelationshipType*. The CDM also introduces element *KnownReference*, to identify the record who triggered a response. As shown in Figure 5, Entity A sends to Entity B an information object with identifier Oficio2222 through iAP. Entity B receives this information and generates a local record in its RMS. As a response to Oficio2222, Entity B produces a new information object and sends it to Entity A as a reply (Oficio3333). To establish this relationship between information objects in the metadata, Entity B indicates in element *KnownReference* Oficio2222 as the object to whom Oficio3333 is a response to.

The results of this comparison helped identify the need to develop a new data model and the reason why MIP was not chosen to be applied directly to the service. MIP provides the elements required in a Portuguese metadata schema to correctly identify the records, but it does not provide a well-formed structure to be applied digitally. MIP was developed focusing more on how to identifying records within an organization and less in how these records would be interpreted if their metadata was shared with a different organization.

4.2 Assessment of the CDM Qualities

The Bruce-Hillman Framework (BHF) is a technique used to assess the qualities of metadata schemas (Bruce and Hillmann, 2004), defining seven qualities: completeness, provenance, accuracy, conformance to expectations, logical consistency and coherence, timeliness and accessibility. Each quality is associated with questions whose answers provide a narrative score, as depicted in Figure 6.

According to the BHF, ***Completeness*** is the capability of the metadata schema to describe the object as completely as possible, considering the project's resources. To measure this quality, the BHF presents two questions. As shown in Figure 6, the first question can be answered affirmatively, considering that

the CDM was developed under the influence of MIP. As concluded in subsection 4.1, MIP was not designed to be applied to records that are meant to be exported to other systems, since it does not ensure that the identity of the record is maintained when exported[11]. However, MIP was developed as a measure for describing records metadata correctly. Even if the way MIP is structured is not ideal and would not generate good results if applied in this context, it is a fundamental reference for CDM. This way, we can sustain CDM supports effectively the creation of the local record, since it follows the requirements from MIP. When developing the CDM, and in similarity with MIP, there was a consciousness that not every element is always required, since many just provide extra information that helps to identify the record, but is not necessary seeing as other elements are capable of providing enough. However, every element that is considered crucial to describe the record is present and is mandatory, meaning that the information provided by those elements is always transmitted, without exception. With this in consideration, the second question proposed by the BHF, shown in Figure 6 is positive, presenting two examples, in column "Compliance indicator", of elements crucial to define the record's identity.

Provenance, as a quality, is defined as the capability the metadata has of providing information about its origins and changes throughout time. To assess that, three questions are proposed, as in Figure 6. The first question tries to understand if the element set provides information about the responsible for creating the metadata. The CDM is capable of providing this information, considering that the responsible for keying the metadata of the record is the generator organization. The next two questions analyze if the metadata provides information of how the metadata was created and if the metadata has suffered any transformations since its creation. The CDM does not provide any information about these issues. The CDM only registers information regarding the version of the record but not changes in its metadata.

Another quality promoted by BHF is ***Accuracy***. Bruce and Hillman state that a metadata schema should be accurate in the way it describes the data object, by providing *"correct and factual"* (Bruce and Hillmann, 2004) information. To assess this, the BHF proposes three questions. As mentioned previously,

[11]The export of a records from one system to another occurs in very specific business scenarios, due for example a legal obligation, for preservation of the original records due to the decommission of an old system, etc. Anyway, even if that is not a specific concern of this work, the results here presented also can contribute for that to be more easily done in the future over the iAP.

Quality Measure	Quality Criteria	Narrative Score	Compliance Indicator
Completeness	Does the element set completely describe the objects?	Yes, the CDM describes the record, providing the required information stated by an approved standard (MIP)	MIP and CDM comparison
	Are all relevant element used for each object?	Yes, all the elements considered relevant and important to ensure the record's identity are mandatory	Elements *Identifier* and *DocumentType*
Provenance	Who is responsible for creating, extracting, or transforming the metadata?	The CDM provides this information through one of its elements	Element *GeneratorOrganizationID*
	How was the metadata created or extracted?	The CDM doesn't provide this information	
	What transformations have been done on the data since its creation?	The CDM is not capable of registering which transformations the data has suffered since its creation	
Accuracy	Have accepted methods been used for creation or extraction?	Yes, the CDM was developed to maintain the elements presented by MIP, while introducing necessary changes to achieve the research's goal	MIP and CDM comparison
	What has been done to ensure valid values and structure?	To ensure this, the work group responsible for the development was composed by different professionals with different visions, to cover al the necessities for valid values and structure	Application profile
	Are default values appropriate, and have they been appropriately used?	*The information available is not enough to answer this question*	
Conformance to expectations	Does metadata describe what it claims to?	Yes, the CDM describes the record as correctly as it claims, since its elements provide information based on MIP	MIP and CDM comparison
	Are controlled vocabularies aligned with the audience characteristics and understanding of the objects?	Yes, when developing controlled vocabularies, users' necessities were considered	Element *DocumentType*
	Are compromises documented and in line with the community expectation?	Yes, all the compromises made during the development phase are documented	Internal documentations; Meeting minutes
Logical consistency and coherence	Is data in elements consistent throughout?	Yes, elements with similar functions have a similar structure throughout the CDM	Elements *XOrganizationType* and *XOrganizationID*
	How does it compare with other data within the community?	The CDM allows a bigger data consistency than MIP	MIP and CDM comparison
Timeliness	Is metadata regularly updated as the resources change?	*The information available is not enough to answer this question*	
	Are controlled vocabularies updated when relevant?	*The information available is not enough to answer this question*	
Accessibility	Is an appropriate element set for audience and community being used?	Yes, considering it provides the same information as MIP, an approved schema within the community	MIP and CDM comparison
	Is it affordable to use and maintain?	Yes, it is affordable due to the flexibility of the language used	Documentation and calculations
	Does it permit further value-adds?	Yes, the CDM allows the temporary addition of elements and due to the language, it was develop in, a permanent element is easily added	Element *SpecificMetadata* and use of XML

Figure 6: Bruce-Hillman framework applied to the CDM (Bruce and Hillmann, 2004).

when developing the CDM, there was a concern in ensuring that the information provided by MIP was maintained by the new elements of the CDM. Considering this, the first question, shown in Figure 6, can be answered affirmatively, seeing as MIP is an accepted method and was used for the creation of the CDM. The second question inquiries about what has been done to ensure valid values and structure for elements of the metadata schema. While developing the CDM, a great deal of importance was given to the structure of the elements, since this was the main factor why MIP was not apt to be used, as shown in subsection 4.1. For defining the CDM, a set of professionals, from different areas within the community, that would bring different ideas and perspectives to the table, were selected to help evaluate which values were needed for each element and what structure the elements should adopt to achieve the goals of the research. Since the CDM has yet to be used by organizations within the PA, the third question cannot be answered due to the lack of information regarding its application.

A metadata schema is in ***Conformance to expectations*** if it is able to respond to the users necessities, by including elements that the community expects to find, while remaining realistic about what is and is not important to be included. To evaluate this quality, BHF proposes three questions. The first, shown in Figure 6, can be answered affirmatively. The defi-

nition of the CDM was driven by the need to guarantee that a record would be described correctly whilst ensuring that this information would not be lose its meaning when received by another system. Considering that CDM elements provide the same information as MIP elements, an accepted standard for defining records within the PPA, it can be said that the CDM describes what it claims. The second question inquiries about the use of controlled vocabularies and if they are aligned with the needs of users and records. When developing the controlled vocabulary for element *DocumentType*, a survey was performed to understand which types of documents were more frequent within the organizations of the PPA. With this information, it was possible to define a controlled vocabulary with 38 types of documents to be used as values for the element. This serves as an example of how users' needs were taken into account when establishing controlled vocabularies for each element. The third question refers to the compromises made throughout the implementation phase, and their registration. To elaborate the CDM, different opinions of archivists and information systems professionals had to be balanced, requiring multiple compromises, documented in meeting minutes and other internal documentation produced during the development phase of the CDM.

Logical consistency and coherence are qualities of a metadata schema with a consistent structure throughout its definition and associated application profile. An application profile is a set of metadata elements with guidelines and policies associated that indicate how the metadata elements are to be applied to the objects. In this case, the CDM proposed is throughly explained in its application profile, exposing to the users the accepted values and utilization norms of every element. As shown in Figure 6, two questions are associated to this quality. The first, questions if the data in elements is consistent throughout. For this assessment, it is possible to evaluate what the CDM offers, to ensure that the data will be consistent throughout. This metadata schema is consistent in the way it describes different elements who have the same functionality in the schema, even if used in different contexts. Elements of type *XOrganizationType* and *XOrganizationID* have the same functionality of identifying organizations, present the same structure and accept the same type of values, but can be applied to different types of organizations and their different roles for the record. Even if there is not, yet, a set of data that can be assessed for its consistency, the CDM has provided all the tools to ensure this consistency. To answer the second question proposed, shown in Figure 6, taking into consideration the lack of application of the CDM, it is only possible to compare the architecture of the CDM with the architecture of MIP, and how consistent the data provided by both data models would be. As stated in subsection 4.1, MIP does not provide rigid rules for the data as the CDM does. Although the MIP provides multiple examples of values to assign to each element, nothing prevents users of applying different rules in the same metadata object. Consequently, this does not ensure that the data will be consistent throughout.

According to the BHF, ***Accessibility*** is the capability a metadata set has to be viewed and comprehended. This quality is assessed by three questions. The first, questions the appropriateness of the element set for the community. As stated previously, MIP was used as the basis for developing the CDM. Considering this, the CDM proposed is appropriate for the community, taking into consideration the concerns from multiple perspectives and providing a solution for them. The second question inquiries about the costs of maintenance of the element set. Considering the use of XML and the results from simulative calculations, it was concluded that maintenance is affordable and even preferable to the costs of correspondence, nowadays, in the PA. The last question audits the ease of adding further elements to the set. The CDM provides element *SpecificMetadata*, which allows the temporary addition of data to the metadata. A more permanent addition to the CDM is also easy to achieve due to the flexibility of XML, the language used for implementing the CDM.

Considering the data available, it was not possible to evaluate the ***Timeliness***, another quality promoted by the BHF, of the CDM since it refers to metadata and controlled vocabularies updates, which have yet to be tested.

5 CONCLUSIONS AND FUTURE WORK

The development of the CDM is the first step to ensure interoperability among records management systems within the PPA. This is a decisive step for achieving interoperability since it guarantees semantic interoperability among organizations, assuring that they "speak the same language" and that the metadata exchanged is descriptive enough, and well structured, to enable its rightful interpretation and the creation of a new record in a new RMS, based on the information provided. By using MIP as the basis for the elaboration of this data model, all the necessary elements for a correct characterization of the records are represented. MIP's vocabulary and structural prob-

lems are addressed in the CDM proposed, which can be considered an improved version of MIP, ready to be applied to information systems with records management capabilities. The results show that the CDM promotes interoperability through iAP, the Portuguese interoperability platform, by ensuring that the information exchanged maintains its original meaning. Results also show that the CDM produced holds 5 qualities, named Completeness, Accuracy, Conformance to expectations, Logical consistency and coherence and Accessibility.

Even though the element set produced relates only with the PPA, this proposal is useful to understand the steps in an information interoperability project, with special attention on the process of establishing a canonical data model to achieve semantic interoperability in a SOA-ESB environment.

Considering that the CDM developed stands for the initial steps of an ambitious project of interoperating all of the different records management systems of the PPA, there are still several steps to be taken to achieve this goal. For the future, it is important that tests are executed, specially an evaluation regarding the level of automation that the system can acquire in the capture of the document and metadata exchanged and storage of new records by the systems, due to the application of the CDM. The CDM may need future improvements according to the feedback provided by the organizations when in use. For example, if the service is highly used to expedite invoices, the addition of element *Price* (*Preço*) may be considered, instead of using the *SpecificMetadata* element, to refer the price associated with the invoice, every time.

ACKNOWLEDGEMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019 and by the European Commission program H2020 under the grant agreement 822404 (project QualiChain). We would also like to show our gratitude for the professionals at AMA and DGLAB who contributed to this research.

REFERENCES

Administrative Modernization Agency (AMA) (2011). Interoperabilidade na Administração Pública, Procedimentos para Adesão à iAP [Interoperability in the Public Administration, iAP Adhesion Guidelines], Version 3.0. in Portuguese.

Barbedo, F. and Corujo, L. (2012). *MIP: Metainformação para Interoperabilidade [MIP: Metadata for Interoperability], Version 1.0c*. General Directorate for Book, Archives and Libraries (DGLAB). in Portuguese.

Bruce, T. R. and Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. ALA editions.

Chappell, D. (2004). *Enterprise Service Bus*. O'Reilly Series. O'Reilly Media, Incorporated.

Dave Hollander (2011). Common Models in SOA. https://statswiki.unece.org/download/attachments/65372409/common_model_in_soa_wp.pdf (Accessed: 15-06-2018).

European Commission (2010). *COM(2010) 744 final, Annex 2 to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions: "Toward interoperability for European public services", European Interoperability Framework (EIF) for European public services*.

European Commission (2017). Improving semantic interoperability in European eGovernment systems. https://ec.europa.eu/isa2/actions/improving-semantic-interoperability-european -egovernment-systems_en (Accessed: 19-12-2017).

European Parliament and of the Council (2009). *Decision No. 922/2009/EC of the European Parliament and of the Council on interoperability solutions for European public administrations (ISA)*.

General Directorate for Book, Archives and Libraries (DGLAB) (2013). *Macroestrutura Funcional (MEF) [Functional Macrostructure (MEF)], Version 2.0*. in Portuguese.

General Directorate for Book, Archives and Libraries (DGLAB) (2014). *Lista Consolidada: 3^{o}s níveis em planos de classificação conformes à MEF ["Lista Consolidada": 3rd levels in business classification schemes according to MEF]*. in Portuguese.

Hohpe, G., Woolf, B., and Brown, K. (2004). *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. A Martin Fowler signature book. Addison-Wesley.

ISO 15489-1:2016 (2016). Information and documentation – Records management - Part 1: Concepts and principles.

Lourenço, A. and Penteado, P. (2015). A caminho da ASIA – Avaliação Suprainstitucional da Informação Arquivística [On the way to ASIA - Suprainstitutional Evaluation of Archivist Information]. In *BAD National Congress*, number 12. in Portuguese.

Lourenço, A., Penteado, P., and Henriques, C. (2012). O desafio da interoperabilidade na gestão dos arquivos da Administração: propostas do órgão de coordenação nacional de arquivos [The challenge of interoperability in Administration's archive management: Proposals from the national coordination archives association]. In *BAD National Congress*, number 11. in Portuguese.

Maguire, R. (2005). Lessons learned from implementing an electronic records management system. *Records Management Journal*, 15(3):150–157. https://doi.org/10.1108/09565690510632337.

Online Computer Library Center (2013). Using a Controlled Vocabulary. https://www.oclc.org/content/dam/training/CONTENTdm/pdf/Tutorials/Metadata/Controlled Vocabulary.pdf (Accessed: 07-08-2018).