24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

A layered meta-data approach to the design of joint experimentation data warehouse

Li Jiang, Maria Boyle, Kevin Johns, Chitra Muniyappagounder Subramanian and Sanat Bista

Strategy and Joint Force, Joint and Operation Analysis Division, (JOAD) Defence Science and TechnologyGroup, Department of Defence
Email: li.jiang@dst.defence.gov.au

Abstract: Force design activities such as Force Structure Planning (FSP), wargaming and experimentation generate large amounts of data in different formats. Current in-house data management systems are limited in their ability to store and manage data efficiently. The development of a Joint Analytics and Reporting System (JARS) as a data warehouse for experimentation activities will provide the necessary infrastructure to support data storage, management and retrieval. Availability of quality, fit-for-purpose data via JARS will also enable the use of advanced analytics techniques to support evidence-based decision-making. Using historical experimentation data to support the planning of future activities is challenging as data sources, formats and data quality are often variable. In this paper, we propose JARS as a practical solution and present a layered metadata model approach based on our research and analysis of the historical data. As part of this endeavour, we conducted *requirements discovery* workshops to assist in the design of JARS. The JARS schema was developed based on database theory, metadata design principles, and best industry practices such as the FAIR principles (Findability, Accessibility, Interoperability, and Reusability of digital assets). Furthermore, we considered the trade-offs between the principles of database design, data integrity, data processing complexity, and completeness and efficiency of querying. The metadata model contains three-layers of information: Campaign layer, Activity layer, and Extensible layer. The Campaign layer contains Campaign name, Campaign objectives, and so forth. The Activity Layer contains metadata information such as Name, Status, Location and Scenario used. The Extensible Layer contains more information and data required for finding nuanced information. This layered structure provides a flexible and effective way to store and query various metadata information as required. It can also assist in achieving a balance between information integrity and data complexity, and between completeness of information and efficiency of data retrieval. The latter is often a challenge in database design. This paper aims to present theresults of our past research and ongoing work. The significance of JARS and the future work are also discussed.

Keywords: *Metadata, data warehouse, force design, design principle, decision support system, experimentation*

1. INTRODUCTION

"*War is ninety percent information*" *- Napoleon Bonaparte.*

Defence regularly conducts core force design activities such as Force Structure Planning (FSP), wargaming, andexperimentation. These activities generate large amounts of complex data characterised by both volume and variety. To handle the growing complexities of data and make the *data-to-insight-to-action* process efficient andeffective, the current in-house data management systems need appropriate modernisation (Jiang *et al.* 2015, Jiang*et al.* 2016). The Joint Analytics and Reporting System (JARS) is designed as a data warehouse for the storage, integration, and interoperability of data. In addition, JARS is an essential infrastructure for the Science Research Computing Environment (SRCE), and the Joint Experiment and Wargaming Laboratory (JEWL) areas within the Joint and Operation Analysis Division (JOAD), Defence Science and Technology Group. JARS will also support data analytics, business intelligence, metadata management, data integrity, quality assurance and security. Addressing the complex data issues occurring across JOAD is very challenging, as data sources and formats used in wargaming and experimentation are varied. We propose JARS as a practical solution and present a layered metadata model based on our research and analysis of the data issues. This paper presents our approach to developing JARS, which includes the process of requirements elicitation and logical and physical database design. We also present our previous research findings, ongoing research, and the future direction of this work.

2. LITERATURE REVIEW

A data warehouse is a "subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decisions" (Inmon and Hackathorn 1994). The terms "subject-oriented" and "integrated" suggest that the data warehouse can provide a comprehensive view of the data to support effective decision-making. The term "time-variant" describes the historical or time-series nature of the data in the warehouse, enabling trend and longitudinal analysis (Inmon and Hackathorn 1994). Developing a high-quality data warehouse is a non-trivial and long-term enterprise (Inmon 2000). There is broad consensus in the literature (for example, Do and Rahm 2000; Jarke et al. 2002; Singh and Singh 2010) that the rapid development of data warehouses at all stages is impeded by the underlying quality of the data and the lack of effective mechanisms to deal with the complexity and heterogeneity of the data at many levels. These challenges of data quality and complexity are also inherent inthe development of JARS.

Analysis of the force assessment and wargaming data over the last decade has revealed that the inclusion of highquality metadata is an important initial step in developing JARS. Metadata is often colloquially known as the 'description of data'. Other research has shown that high-quality metadata increases user confidence in the data (Foshay *et al.* 2014). Notwithstanding, there is a general lack of reliable and scalable methods for managing andprocessing data (Greenberg 2005, Gonçalves *et al.* 2017). Typically, the existing mechanism to deal with the metadata is very limited. Metadata models are not common in science (Schembera and Iglezakis 2018) despite their essential role in applying FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles(Wilkinson *et al.* 2016). The FAIR data principles are a guideline for the best practices in database and data warehouse development and data management.

Schembera and Iglezakis (2018) emphasised that well-structured metadata needs to address three requirements: to find the data, to understand the data and to replicate the data. A very well documented metadata model is the Dublin Core Metadata Element Set (Dublin Core), used for searching library resources. The Dublin Core is limited to finding basic search fields such as author, format, language, ID number, and title (Weibel *et al.* 1998). Based on the Dublin Core, the Data Catalog Vocabulary (DCAT) was developed (Maali *et al.* 2010). DCAT was designed to increase interoperability between different data catalogues and simplify the search for data sources (Reiche and Höfig 2013). However, both the Dublin Core and DCAT lack structure and are limited in their ability to incorporate both generic and specific scientific metadata components, while simultaneously being able to find highly detailed information (Matthews *et al.* 2010). To address these issues, a range of models has been developed, and a description of some of these follows.

The Core Scientific Meta-Data (CSMD) model was designed to capture a common set of information about the data produced by experiments, measurements and simulations in laboratory sciences (Matthews *et al.* 2010). CSMD allows users to capture common features of the data. These include descriptions of the data production process (e.g. ID, date, time, etc.); the format, data type, owner of the data; the data parameters and the relationships between the data. Gonçalves *et al.* (2017) presented a workbench developed to improve the quality of metadata submitted to scientific data repositories in the

Center for Expanded Data Annotation and Retrieval(CEDAR) (Gonçalves *et al.* 2017). The CEDAR workbench is a set of open-source, web-based tools for the acquisition, storage, search and reuse of metadata templates, and enables the user to construct templates for metadata. The metadata produced with CEDAR templates conform to the FAIR data principles and are interoperable with Linked Open Data. CEDAR metadata are retrievable in JSON, JSON-LD, and RDF formats (Gonçalves *et al.* 2017).

Schembera and Iglezakis developed a very specific metadata model for research data in computational engineering (Schembera and Iglezakis 2018). At the centre of the model is the data set composed of several files, described in terms of file type, generic metadata such as description, keywords, classification of project, and so forth. The focus of the model is on domain-specific metadata categories with an entity processing step delivering information about the research and describing the methods and the software and hardware used (Schembera and Iglezakis 2018, Schembera and Iglezakis 2020). The most relevant metadata model we found is the Common European Research Information Format (CERIF) model, which is a three-layered data model, maintained by the euroCRIS community (Jeffery *et al.* 2014). The CERIF model contains the following three layers:

- the discovery layer which is used to find data that may be of interest
- the context layers which is a superset of discovery metadata and defines applicable relevant information for associated underlying datasets, and
- the detailed layer that is specific to the data it describes, such as a particular set of data or to a domain producing multiple datasets in a standard format.

In our research, we found that there is no one model that provides a perfect-fit solution for our problem despite the models discussed above having merits in certain ways. Collectively, however, they suggest useful insight forour JARS developmental work, which is discussed in the next section.

3. JARS DESIGN METHODOLOGY – A METADATA DRIVEN APPROACH

Research suggests that a data warehouse should be built iteratively (Inmon 2000). The common practice is to quickly develop a small implementation of the highly valued functionality of the data warehouse by studying a set of core data requirements that addresses the critical business needs. Based on this principle and our experience in database design, we took the following systematic approach in the research and design of JARS. The steps taken in this approach are:

1. Conduct research into discovering the key issues and requirements through a stakeholder requirementsdiscovery workshop.
2. Analyse requirements from the workshop, followed by a synthesis and prioritisation of the requirements.
3. Study the meta-data design methodologies and explore the best solution for meta-data design.
4. Design the metadata logic and physical database schema based on Step 2 and 3 above.

3.1. Requirements elicitation and validation

Based on our experiences with various force design activities and research over the past 10 years, we developed a set of data warehouse requirements (Jiang *et al.* 2015, Jiang *et al.* 2016, Jiang and Bista 2018) essential for building an effective data solution to support force design, wargaming and experimentation activities. In order to maximise JARS's utility for JOAD analysts and Defence users, a requirements discovery workshop was held during the problem definition phase. Representatives with considerable experience in data collection, processing, and analysis across JOAD participated. Through the workshop, we developed a common understanding of the key issues and data requirements related to JARS. We categorised and prioritised the requirements identified in the workshop (Table 1), and developed options for creating appropriate solutions to best support our major tasks.

We found that the majority of the requirements identified in the workshop are consistent with our previous research findings (Jiang *et al.* 2015, Jiang *et al.* 2016). That is, the results strengthen the necessity of improving our data management practices within the organisation and providing effective support for advanced analytics and decision-making.

Table 1. Outcomes of the requirement discovery workshop

Issues	Issue Description
Data collection & quality	Data comes from multiple disparate sources and are predominantly unstructured. Application of quality control mechanisms is difficult.
Data storage & searchability	Data is stored under various locations that are not easily accessible. With the given storage and formatting complexities, data is not searchable.
Data attributes - context and metadata	Data collection and presentation includes limited information on the context of the collection. Use of technical jargons adds to the ambiguity and makes data capturing and interpretation harder.

Data format and structure	Data files have various formats: e.g. text, numbers, digital, excel, .doc, pdf, video, audio. Data is often embedded deep within the files, inside paragraphs of text, tables and figures, and hard to extract.
Historical data and repeated experiments	o Historical data that are required to support longitudinal analysis is not easily accessible, causing repetition of similar experiments at times.
Data integration	Raw data is seldom integrated. Experimentation activities use different approaches and generate different kinds of data sets. Making a general data model is a challenge.
People – Access and User Needs	There are many competing deadlines and there may be insufficient time to process large amounts of data. Data solutions should be beneficial for all users: “a useful solution for analysts with different needs”.
Strategy and Reporting	Need to be able to aggregate data into reports. Data must be interpretable for force design and strategic decisions.

3.2. Layered Meta-data model

Developing a data model is essential in building JARS, as this will provide a consistent source of information foranalysts and users across Defence. Wargaming and experimentation activities tend to be problematic in that they use different approaches, terminologies, and models that make it difficult to create a single data model for all applications and generate consistent data. In the early stage of the JARS development, we investigated various existing data warehouse solutions (Bontempo and Zagelow 1998, Gardner 1998, Jones *et al.* 2001, Paim and de Castro 2003, Ariyachandra and Watson 2005, Ferreira and Furtado 2013). We took a hybrid approach that utilised both top-down (Inmon 2005), and bottom-up (Moody and Kortink 2000) methods. Based on the analysis of the available data and data model in force design, we found that a metadata-oriented model could provide an effective solution at the early stage of the project. We used both metadata design principles (Duval *et al.* 2002, Greenberg 2005) and the FAIR principles. The FAIR principles emphasise machine-actionability, that is, the capacity of software applications to find, access, interoperate and re-use data with minimal human intervention. Meta-data design principles encourage using namespaces and metadata modularity, improving extensibility, and the use of an iterative refinement approach while considering the practicalities of the design and implementation. This alsoincludes developing a metadata association model, using metadata registries, evaluating mandatory versus optional elements, and subjective and objective metadata. The metadata design for the JARS should be based onthe physical structure and characteristics of the original data. Using FAIR and metadata development methodologies discussed above, a set of functionalities with higher priority for the first stage of JARS implementation was developed as follows:

- Metadata storage and processing for effective contextual data management.
- Flexible data format & structure management for effective data query/access
- Data curation and data integration for supporting data from various sources.

Based on this set of functionalities and the metadata design principles discussed above, we developed a layered metadata model, which allows stakeholders to store, easily find and access experimentation data through search queries at various levels of granularity and detail (Fig. 1). The metadata model contains, minimally, three layers: Campaign Layer, Activity Layer, and Extensible Layer, which is able to be further extended if necessary. The Campaign Layer (the first layer) contains the essential information about the campaign such as campaign name and objectives, and is the atomic level information for a campaign. A campaign can contain many activities that are located at the second layer–the Activity Layer. This layer contains the metadata information about an activity such as Name, Status, Location, Scenario used, and so forth. Some data items in this layer are atomic such as Activity Name; many data items are composite which are also extensible (meaning more information can be injected into the database as required if available). For instance, a scenario of an activity can be simply saved in the scenario description column, while it can also be saved in multiple files that contain detailed information about the scenario such as graphs, links to the strategic objective, and/or military strategies. Likewise, the data item, Activity File, can contain simple information such as a range of links to the physical files, and can contain detailed information about the file types, which can be Word, PDF, Excel, and image files such as JPG, and so forth.

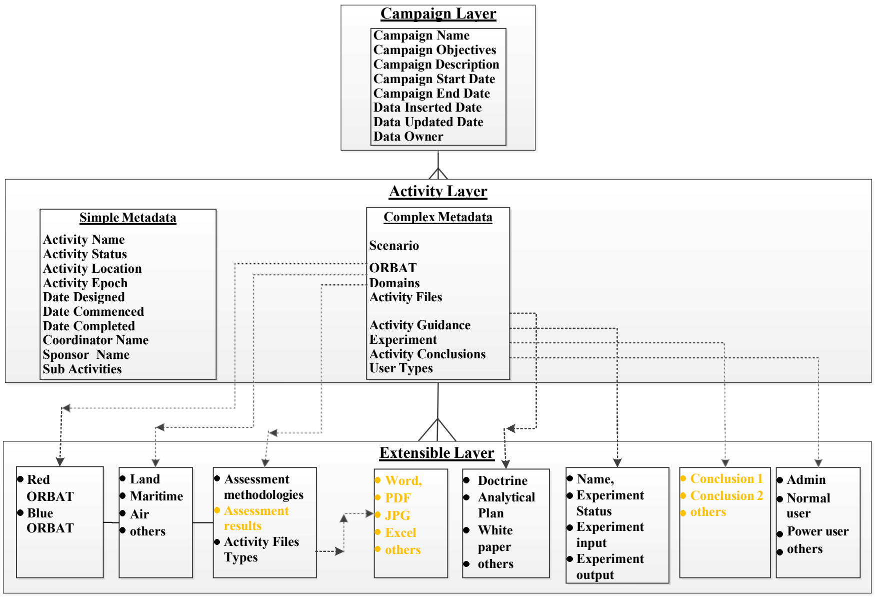

Fig. 1 Metadata Model

Legend: "___<" indicates one to many relationship; for instance, at Campaign level, a campaign might contain one or many activities."
---▶ " represents the association between the metadata.
The metadata in orange represent outputs from the activities, whilst, the metadata in black represent input data.

This layered structure provides a flexible and effective way to store and query various metadata information and finddata files.

The model aims at providing an effective solution for the following problems:

- metadata collection: storing, organising, and searching the data through a metadata inquiry;
- managing metadata and meta information system with the functions of identifying, searching, retrieving,processing, analysing, and interpreting; and
- developing, managing and evolving metadata models.

The layered metadata model and data scheme developed above are expected to be used as a framework for a moreconsistent taxonomy of the data acquisition requirements within the Force Experimentation process, which can provide a more structured data approach within the future JARS Data Warehouse. Based on our analysis and the dry manual test of the proposed model, we found that the proposed metadata model could provide a flexible way to store and process metadata at a different level of modularity and granularity. A physical data model was also developed based on the metadata model. Our proposed schema, once implemented in the physical model in the JARS data warehouse, will provide an effective mechanism for data input, storage, and management. JARS is expected to provide a single source of the truth for data management across the entire SRCE network in the nearfuture.

3.3. The Utility of the Metadata Model

Fig. 2 presents an entity relationship diagram between tables at the implementation level only. We will present several data query examples below to illustrate the utility of the metadata model. In the following table, we assume that the force assessment data including their metadata over the last two decades were stored in the JARS data warehouse. Table 2 shows data and information that can be extracted from the database. The campaign layer hashigher-level information namely: campaign name, campaign objectives, campaign starting time, campaign ending time, etc. The campaign objectives contain information about the aims of the campaign or the intent of the commander at the highest level. For instance, the campaign objective might be "to examine the effectiveness of ADF Electronic Warfare in Defending Wonderland Island (an imaginary country)". The activity layer contains details of campaign activities such as activity name, sponsor for the activity, activity status, commencement andcompleted date for an activity, sub-activities and so forth. The extension layer contains nuanced detailed information about the campaign and activities. Some data items are beyond metadata, which contains essential information often required by DSTG analysts. Some examples of the questions and corresponding SQL queries that can be used in the three-layered metadata database are shown in Table 2. Besides the exemplar utilities shownin Table 2, more complex inquiries can be made. For example:

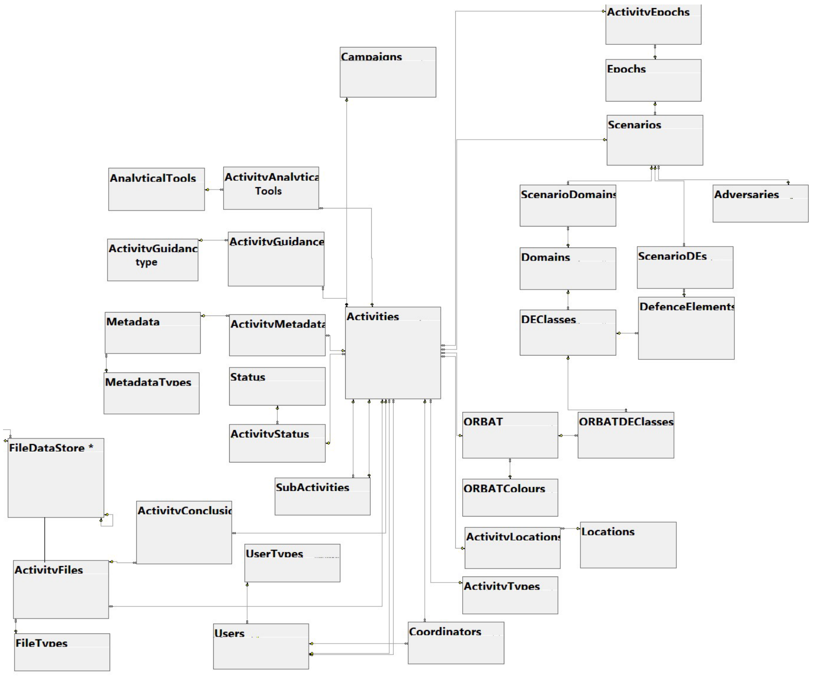

Figure 2. Metadata Entity Relationship Diagram

Table 2. Query examples of three layered metadata database

Layer	No.	Query Example	Expected Result
mpaignlayer	1.	What are the objectives of a given campaign **Y**?	The given campaign objectives and some additional campaign description
	2.	What is the start and end date for a given campaign **Y**?	Campaign start data, campaign end date
	3.	When was the campaign data latest updated and who did the update for a given campaign **Y**?	The date of the campaign data being inserted into the database and the person's name who made the update on the campaign data.
ctivitylayer	1.	What are the names of the past campaigns and their activities?	A list of all past campaign names, their activities and the sub activities contained in each activity.
	2.	Who is the sponsor for a given activity **X**?	A list of Sponsor Names and Activity Names
	3.	Who is the designer for a given activity **X**?	The designer's name of a given activity.
	4.	Who is the coordinator for a given activity **X**?	The coordinator's name of a given activity.
	5	What are the major conclusions of activity **X** in and their risks in Campaign **Y**?	A list of conclusions and their risks in Activity X of CampaignY
tensionlayer	1.	What analytical tools are used for activity **X**?	The activity name, analytical tools used in the activity and the description of the tools for the given activity X
	2.	Where (physical location of) is the analytical plan document for activity **X?**	The activity name, and its analytical plan document's physical location, and the document description for activity X

- What are the ORBATs for blue or red force for a given scenario in a given activity? This type of query will involve more than eight tables in our metadata database that are not included in the above example for simplicity.
- What sort of campaign or wargaming activities have we done over the last 10 years with the "Effectiveness of Electronic Warfare" as the objective? These sort of questions are more complicated than all types of queries illustrated above. Consequently, the solution requires utilising combined approaches. For instance, we can usetext searching/mining algorithms to find the best match of the phrases "Effectiveness of Electronic Warfare" or "Electronic Warfare" in the Campaign Objective and Activity Objective columns in the corresponding metadata tables.
- What are the persistent issues or gaps in ADF capability that can be found in the campaigns and wargaming activities over the last 10 years? To find such an answer is not only critical for analysis, but also possible withthe support of the metadata database using advanced machine learning and artificial intelligence (AI) techniques.

4. CONCLUSION AND FUTURE WORK

Developing a fully effective JARS data warehouse to support FSP and associated wargaming activities is a long-term endeavour, which will be built through an iterative development process (Inmon 2000).
We have shown that a layered metadata driven approach can provide the following advantages:

- For end-users to input data in a flexible way, as data become available
- To store various types of data related to force design activities
- To input data at multiple levels, (campaign level, activities level, and extension level) which allows tailored storage and retrieval when necessary
  - To conduct longitudinal study of historical data to gain critical insight into ADF capabilities.

Our future work will focus on the following areas:

- Continuous refinement of the metadata model and populating data into JARS on a regular basis
- Develop a web application front-end that can be accessed easily and offers on-demand analytical features
- Research, develop and implement AI technologies to enhance discoverability and analysis of JARS data holdings.

ACKNOWLEDGEMENT

Our sincere thanks to all participants of the JARS requirements discovery workshop. Special thanks to Cayt Rowe, Peter Hoek, Ed Lo, Andrew Au and James Peacock for useful conversations and feedback over the architecture and design of JARS.

REFERENCES

Ariyachandra, T. and H. J. Watson (2005). "Key factors in selecting a Data Warehouse architecture." Business Intelligence Journal **10**(2): 19-26.

Bontempo, C. and G. Zagelow (1998). "The IBM data warehouse architecture." Communications of the ACM **41**(9): 38-48.

Duval, E., W. Hodgins, S. Sutton and S. L. Weibel (2002). "Metadata principles and practicalities." D-lib Magazine **8**(4): 1-10.

Ferreira, N. and P. Furtado (2013). "Real-time data warehouse: a solution and evaluation." International Journalof Business Intelligence and Data Mining **8**(3): 244-263.

Foshay, N., A. Taylor and A. Mukherjee (2014). "Winning the hearts and minds of business intelligence users:The role of metadata." Information systems management **31**(2): 167-180.

Gardner, S. R. (1998). "Building the data warehouse." Communications of the ACM **41**(9): 52-60.

Gonçalves, R. S., M. J. O'Connor, M. Martínez-Romero, A. L. Egyedi, D. Willrett, J. Graybeal and M. A. Musen (2017). The CEDAR workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments. International Semantic Web Conference, Springer.

Greenberg, J. (2005). "Understanding metadata and metadata schemes." Cataloging & classification quarterly **40**(3-4): 17-36.

Inmon, W.H. (2000). "Building the data warehouse: Getting started." WebLink: http://www.academia.edu/3081161/Building_the_data_warehouse.

Inmon, W. H. (2005). Building the data warehouse, John wiley & sons.

Inmon, W. H. and R. D. Hackathorn (1994). Using the data warehouse, Wiley-QED Publishing.

Jeffery, K. G., A. Asserson, N. Houssos, V. Brasse and B. Jörg (2014). "From open data to data-intensive science through CERIF." Procedia Computer Science **33**: 191-198.

Jiang, L. and S. Bista (2018). Towards An Enduring Force Design Decision Supporting System, presented in. Defence Operations Research Symposium. Melbourne, Australia.

Jiang, L., N. Tay and G. Bulluss (2016). Exploring a feasible data architecture for supporting Force Design. Defence Operations Research Symposium. Canberra.

Jiang, L., N. Tay, H. Seif Zadeh and G. Bulluss (2015). Towards Defence Strategic Data Planning. 21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015 A. Gold Coast. Gold Coast, Australia,.

Jones, M. B., C. Berkley, J. Bojilova and M. Schildhauer (2001). "Managing scientific metadata." IEEE Internet Computing **5**(5): 59-68.

Maali, F., R. Cyganiak and V. Peristeras (2010). Enabling interoperability of government data catalogues. International Conference on Electronic Government, Lausanne, Switzerland, Springer.

Matthews, B., S. Sufi, D. Flannery, L. Lerusse, T. Griffin, M. Gleaves and K. Kleese (2010). "Using a core scientific metadata model in large-scale facilities." The International Journal of Digital Curation **5**(1).

Moody, D. L. and M. A. Kortink (2000). From enterprise models to dimensional models: a methodology for data warehouse and data mart design. The Second International Workshop on Design and Management of Data Warehouses, Stockholm, Sweden.

Paim, F. R. S. and J. F. B. de Castro (2003). DWARF: An approach for requirements definition and managementof data warehouse systems. Proceedings of 11th IEEE International Requirements Engineering Conference, 2003., Monterey Bay, CA, USA, IEEE.

Reiche, K. J. and E. Höfig (2013). Implementation of metadata quality metrics and application on public government data. 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, IEEE. Schembera, B. and D. Iglezakis (2018). The Genesis of EngMeta-A Metadata Model for Research Data in Computational Engineering. Research Conference on Metadata and Semantics Research, Springer.

Schembera, B. and D. Iglezakis (2020). "EngMeta: Metadata for computational engineering." International Journal of Metadata, Semantics and Ontologies **14**(1): 26-38.

Weibel, S., J. Kunze, C. Lagoze and M. Wolf (1998). "Dublin core metadata for resource discovery." Internet Engineering Task Force RFC **2413**(222): 132.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos and P. E. Bourne (2016). "The FAIR Guiding Principles for scientific data managementand stewardship." Scientific data **3**(1): 1-9.