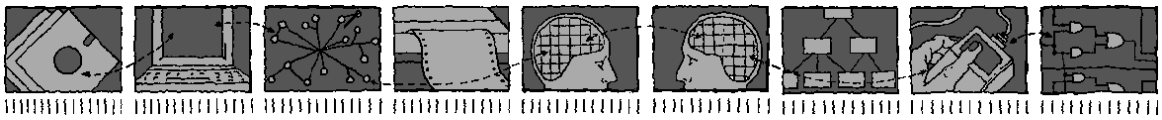
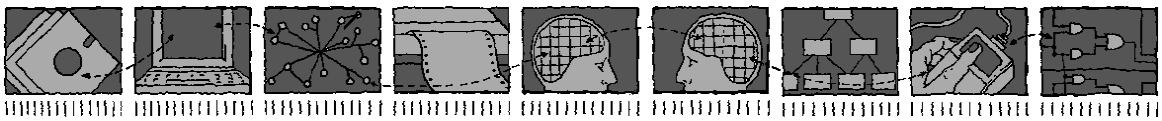
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Stirling Online Research Repository



*Department of Computing Science and Mathematics*
*University of Stirling*



# The DAMES Metadata Approach

**Jesse M. Blum, Kenneth J. Turner**

*Technical Report CSM-177*

*ISSN 1460-9673*

December 2008

*Department of Computing Science and Mathematics*
*University of Stirling*

# The DAMES Metadata Approach

**Jesse M. Blum, Kenneth J. Turner**

Department of Computing Science and Mathematics
University of Stirling
Stirling FK9 4LA, Scotland

Telephone +44-786-467421, Facsimile +44-786-464551
Email {jmb, kjt}@cs.stir.ac.uk

*Technical Report CSM-177*

*ISSN 1460-9673*

December 2008

# Abstract

The DAMES project will provide high quality data management activities services to the social science research community based on an e-social science infrastructure. The infrastructure is supported by the collection and use of metadata to describe datasets and other social science resources. This report reviews the metadata requirements of the DAMES services, reviews a number of metadata standards, and discusses how the selected standards can be used to supprot the DAMES services. The kinds of metadata focussed upon in this report include metadata for describing social science microdatasets and other resources such as data analysis processing instruction files, metadata for grouping and linking datasets, and metadata for describing the provenance of data as it is transformed through analytical procedures. The social science metadata standards reviewed include:

- The Common Warehouse Metamodel (CWM)
- The Data Documentation Initiative (DDI) versions 2 and 3
- Dublin Core
- Encoded Archival Description (EAD)
- e-Government Metadata Standard (e-GMS)
- ELSST and HASSET
- MAchine-Readable Cataloging (MARC)
- Metadata Encoding and Transmission Standard (METS)
- MetaDater
- Open Archives Initiative (OAI)
- Open Archival Information System (OAIS)
- Statistical Data and Metadata Exchange (SDMX)
- Text Encoding Initiative (TEI)

The review concludes that the DDI standard version 3.0 is the most appropriate one to be used in the DAMES project and explains how best to integrate the standard into the project. This includes a description of how to capture metadata upon resource registration, upgrade the metadata from accessible resources available throughthe GEODE project, use the metadata for resource discovery, and generate provenance metadata during data transformation procedures. In addition, a "metadata wizard" is described to help with data management activities.



# Acknowledgements

Funding for The DAMES project is provided by the UK Economic and Social Research Council (grant RES-149-25-1066). The authors wish to thank their colleagues at the University of Stirling and the University of Glasgow for their collaboration.

# 1 Introduction

An enormous amount of statistical data is generated, collected, analysed and stored by governments, institutions, organisations and researchers around the world performing microsocial analysis. The data can consist of different types of resources such as statistical data resulting from microdata surveys, administrative information, synthetic datasets such as time series macrodata, and other data resources such as instructions for aggregating datasets. Data describing the resources, called metadata, can provide considerable value to social science researchers. In general, metadata can be used to document bibliographic information about the resources, provide historical and other contextual details about the data, describe resource relationships and aggregation instructions that can be used to link and compare resources, and support archiving and resource access procedures.

In order to improve data management and exchange, specialised metadata standards have been designed to describe social science resources. In the early years of social science research, metadata was recorded by data producers in an ad hoc fashion using a variety of non-standard techniques. Whilst ad hoc metadata production continues, standardised metadata formats have facilitated a research environment that allows greater data discovery and access, collaboration between researchers from a variety of institutions, and improved data processing capabilities. Standard metadata descriptions enhance social science practices such as resource reuse and transformation, replicating study results, computing and analytical package interoperability, respecting confidentiality, and understanding data quality, relevance, and integrity.

The project Data Management through e-Social Science (DAMES) is a National Centre for e-Social Science (NCeSS) node that aims to develop e-infrastructure for quantitative data analysis of data on ethnicity, educational qualifications, occupations, health and social care. The DAMES project team has considered appropriate metatadata standards to describe heterogeneous data resources that are planned to be made available through DAMES services.

This report describes the DAMES metadata requirements in section 2. A survey of metadata standards for social science is given in section 3. The standard recommended for use in the DAMES project is presented in section 4, along with a description of future work integrating the standard into the project.

# 2 DAMES Requirements

A primary goal of the DAMES project is the design of a grid-enabled framework that incorporates a metadata element profile supportive of the rest of the project's activities. The metadata standard selected for use in the DAMES project has to meet three criteria. Firstly, the standard must be appropriate for the resources exploited in DAMES. Secondly, the standard must support the discovery and usage of resources available through a prior project called GEODE. Thirdly, the standard must ensure compatibility with existing metadata standards used in social science. The first and the second of these criteria are examined below, whilst the third has informed the metadata standards survey presented in the section 3.

## 2.1 DAMES Resources

The DAMES project aims to facilitate high quality empirical research into the domains of occupation, education, ethnicity, health and social care, through data management of microdatasets. The data management activities will concentrate on providing services for variable operationalisation and linking related datasets.

Variable operationalisation involves mapping dataset variables onto summary indicators. Summary indicators are widely-accepted encodings of particular variables. The Standardised Occupational Index is an example of a occupational data summary indicator. Generating mappings between dataset variables and summary indicators involves specialised knowledge about the datasets and acceptable harmonisation procedures. Procedures for recoding data are central to the processes of variable operationalisation. Recoding data in a consistent manner enhances the transparency and replicability of the results. Mapping resources are often represented as statistical analysis software package instruction files (such as Stata Do files). One of the goals of DAMES is to facilitate the discovery and usage of published mapping resources developed by social scientists skilled in the complex processes involved.

Linking datasets from related domains is another of the goals of DAMES. Data linkage is generally founded on the identification of common or comparable variables in different datasets. Variable operationalisation may have to be performed on one or more of the linked datasets during the linkage process. There are two types of grouped datasets that the DAMES services will support: comparison-by-design grouped datasets and ad hoc compared datasets. Comparison-by-design datasets are those which are established as a series of surveys taken over time, such as longitudinal studies with a number of "waves" of study. It is common for the waves of study to evolve over time and describing the changes is an important factor in grouping the waves. Ad hoc compared datasets, such as linked demographic trends and survey data fused with census data, are independent datasets with comparable variables, but no direct connection to each other.

Services for variable operationalisation and linking datasets will require controlled procedures to access, register, deposit, analyse, process and retrieve resources. The primary resources that will be used in DAMES include microdatasets, stored metadata, and processing instructions. The microdatasets include social survey microdatasets about demographic trends and service use, datasets from domestic monitoring of activities of daily living, administrative datasets, census data, microsimulation processing instructions, and processing workflow descriptions. In addition, DAMES will also support resources describing access and authorisation policies, geo-spatial information, and workflows. Some of these types of resources will be deposited into a DAMES repository. Others will be accessed remotely, but be registered with DAMES. Yet others will be produced on the fly by DAMES.

## 2.2 The GEODE Project

The DAMES project is being informed by the Grid Enabled Occupational Data Environment project (GEODE) [2], [3]. GEODE demonstrated the use of metadata schemas to describe occupational data resources. GEODE primarily used the metadata for discovering similar resources, an important step upon which DAMES can build.

GEODE facilitated access to and statistical analysis of occupational datasets through a grid computing environment. It provided a web portal to discover and access registered and deposited datasets, and use analysis services. GEODE created facilities for indexing and curating datasets, and controlled their access. GEODE's grid and portal framework was based on Gridsphere [5], Globus toolkit version 4 [5], and OGSA-DAI [6]. GEODE offers DAMES with a range of occupational datasets and architectural models to expand.

The DAMES project is examining the use of an alternative metadata model to the one used by GEODE. A modified version of the Data Documentation Initiative (DDI) version 2 [4] was used in GEODE to annotate datasets with additional social science information metadata. The metadata was used for resource discovery and resource access. However, the DDI was modified to accommodate a subset of the standard and to support additional types of resources than the survey datasets that DDI 2 was designed to support. This extension of DDI was less than optimal because it inhibited reuse of the existing standard and is not necessarily appropriate for use with a variety of social science resources. As a result, the DAMES project is examining alternative ways of describing the various resources, subsetting standards in a consistent manner, and analysing whether additional functionality can be achieved through creative use of metadata.

DAMES will build upon the successful work undertaken on GEODE. DAMES aims to provide enhanced facilities for depositing new occupational information resources to further populate the GEODE repository. Furthermore, DAMES is planning to provide additional analytical queries for the GEODE stored resources. A chief consideration will be to either support GEODE's existing metadata profile, or to migrate the metadata for GEODE's deposited resources to the profile chosen for DAMES.

# 3 Metadata Standards Survey

DAMES should ensure compatibility with existing metadata standards or ontologies used in social science research. This section is a survey of the capabilities, benefits and disadvantages of social science metadata and ontology standards that could be used for DAMES.

## 3.1 Common Warehouse Metamodel (CWM)

Common Warehouse Metamodel (CWM) was designed by the Object Management Group (OMG) [8] to model objects stored in data warehouses and to control their exchange. Version 1.1 is the current standard. CWM primarily focuses on the exchange of business intelligence metadata between repositories [9]. UML notation is used to represent metamodels and to describe object semantics. CORBA is used to manipulate the objects, and the OMG metadata interchange standard XML Metadata Interchange (XMI) is used to exchange metadata. XMI is specified for the transformation of CWM metamodels into CWM Document Type Definitions (DTD), and ensuring transfer conformant XML documents.

It is the opinion of the authors that CWM is more suitable for business users than social scientists. Being tightly integrated with CORBA makes this standard less appropriate for the grid computing environment that DAMES is creating.

## 3.2 Data Documentation Initiative

The Data Documentation Initiative (DDI), produced by the DDI Alliance (a consortium of 25 institutions from North America and Europe), is an XML standard to describe social science microsurvey datasets and resources. There are three versions of the DDI.

The first version, published in 2000, focused on archiving and documenting completed single surveys. DDI 2 extended the model of DDI 1 by including support for aggregated datasets, in the form of multidimensional tables, and geographic material. Both versions were based on single DTDs that described single data files and were designed for end-user data discovery. Data producers tended to treat description with these versions of DDI as added costs in the data collection process. No mechanisms were natively included in these versions of DDI to extend and adapt the standard for specific needs. As such, projects like GEODE that described unsupported resources (such as files containing summary indicators and variable mappings) needed to do so using non-standard extensions.

DDI 3.0, published in April 2008, uses an entirely new "data lifecycle" model [11]. It contains a larger set of schemas that cover significantly more use cases. DDI 3.0 has much greater concern for reusing content and machine processing. For the DAMES project, there are key improvements over the previous versions:

- Support for additional resource types beyond studies
- Schemas for grouping and comparing resources
- DDI 3.0 was designed to work with other metadata standards including Dublin Core, ISO 19115 (for geographic annotation), METS, OAI and SDMX [1]

Take-up of DDI 3 is not without risk and criticism. Bill Bradley of Human Resources and Social Development Canada described a number of key concerns in [10]. Some of these include lack of tool support (especially NESSTAR), unclear audience (the first versions were clearly targeted at end users, whereas version 3 attempts to change data production procedures as well as provide support for end user data consumption), and the invisibility of content. These concerns to some degree will be mitigated by DAMES. Ideally, DAMES will deliver tool support for metadata collection and processing along with a "killer app" to help support data visibility. For DAMES there are additional concerns. Firstly, the set of schemas describing the standard is significantly more complex than the previous version. This can be alleviated to a degree because the standard has mechanisms to create subsets of the schemas as profiles, and also has native support for standard extension. Secondly, since the standard is new, few resources have been described yet using it. DAMES will need to support metadata described using both versions of DDI, or provide an upgrade path to re-describe DDI 2 resources with DDI 3. Whilst DDI 3 take-up is challenging, the associated benefits of using it compared to other standards makes it attactive.

[1] In designing DDI 3, the authors reviewed the benefits and disadvantages of various metadata standards in [23].

### 3.3 Dublin Core

The Dublin Core metadata element set [12] is used for cross-domain resource description. It is an ISO standard (ISO 15836) that uses of the Resource Description Framework (RDF) to make items easier to search for. The simple level of the standard includes 15 elements while the qualified level adds an additional three as well as mechanisms to refine the semantics. Dublin Core is too simplistic to be of great use in the DAMES project, although support for resources already described using it could be advantageous. Other metadata standards in social science (such as DDI 3) have been designed to support Dublin Core because of its widespread adoption.

### 3.4 Encoded Archival Description (EAD)

The Encoded Archival Description (EAD) is a DTD-based XML standard for archiving corporate records and personal documents [13]. It is maintained by the US Library of Congress in partnership with the Society of American Archivists. The latest version (2002) also supports an XML schema. The standard describes entire documents and is not appropriate for fine-grained variable descriptions, as is required by the DAMES project.

### 3.5 e-Government Metadata Standard (e-GMS)

The e-Government Metadata Standard (e-GMS) facilitates UK public sector resource exchange [14]. The standard is based on Dublin Core. It describes whole resources, although it does contain an aggregate tag, but this is insufficiently fine-grained to be of relevance to DAMES.

### 3.6 MAchine-Readable Cataloging (MARC)

MAchine-Readable Cataloging (MARC) [15] [17] is a standard for representing bibliographic information and is primarily used for library catalog record exchange. The MARC record structure is defined in ISO 2709. MARC offers a number of fields to annotate resources, and there is a mapping between DDI and MARC.

### 3.7 Metadata Encoding and Transmission Standard (METS)

The Metadata Encoding and Transmission Standard (METS) [18] is an open, non-proprietary, XML common object format standard to manage deposited resources in a repository, and to handle their exchange. METS was developed by the US Digital Library Federation (DLF) and is supported by the US Library of Congress. Two of the most common repository systems (DBase and Fedora) support METS. METS supports a flexible approach to descriptive, administrative and structural metadata for any type of digital object [1]. METS has a large user base including the UKDA, the US Library of Congress, and a number of university libraries. As pointed out in [1], METS' use of namespaces across DTDs is inconsistent, and may lead to long-term resource inaccessibility.

### 3.8 MetaDater

The MetaDater project ran from 2003 to the end of 2005 and had the objectives of developing large-scale comparative survey metadata standards, as well as tools for principal investigators and data providers to support the standards. DDI version 3.0 now has support for this type of metadata, thereby making this standard redundant.

### 3.9 Open Archives Initiative (OAI)

The Open Archives Initiative (OAI) [19] provides a framework for federating E-Print and other archive repositories. The framework includes a protocol for metadata harvesting (OAI-PMH) that content providing repositories can use to interface metadata. It is a simpler, but less complete solution than the ANSI/ISO Z39.50 standard (ISO 23950). Metadata from OAI data providers must be specified using the unqualified Dublin Core Metadata Element Set. The OAI has been used in grid and portal projects such as the German Collaborative Climate Community Data and Processing Grid (C3-Grid). C3-Grid uses a

portal that extends OAI to allow metadata in any XML format (not just Dublin Core) to be harvested and searched using Apache Lucene [20]. Although it was shown that OAI is not necessarily limited to Dublin Core, it still has not been shown that OAI can support fine-grained elements of social science datasets as well as can more specific standards such as DDI.

### 3.10 Open Archival Information System (OAIS)

The Open Archival Information System (OAIS) [21] is an ISO reference model representing a long-term archival organisation and is not a metadata standard. It recognises roles for data producers and consumers as well as archive management. Packaged data objects are represented, stored and exchanged. Different packages support different use cases, For instance, Submission Information Packages (SIP) are sent to archives from producers. The OAIS reference model is used by the UKDA [1].

### 3.11 Statistical Data and Metadata Exchange (SDMX)

Statistical Data and Metadata Exchange (SDMX) 1.0 is an ISO standard (ISO 17369) meta-model for exchanging large scale statistical information amongst international organisations. It is composed of different formats for different types of datasets (such as time series) and metadata sets [22]. SDMX mainly focuses on the exchange of aggregated datasets, expecially time series. SDMX 2.0 provides dataset registry interfaces. SDMX and DDI 3.0 are complementary standards and have been designed with alignment in mind. The two standards can overlap in that SDMX can represent microdata, and DDI can represent aggregated data, but the XML of each is specialised such that SDMX is macrodata-focussed, while DDI is microdata-focussed.

### 3.12 Social Science Thesaurii

The European Language Social Science Thesaurus (ELSST) [16] is a multilingual social science thesaurus produced by the UK Data Archive (UKDA) based on their Humanities and Social Science Electronic Thesaurus (HASSET). There are multiple versions of the ELSST, including an RDF version, to support interoperability with metadata standards and other thesaurii. The ELSST was checked against the Council of European Social Science Data Archives (CESSDA) and UKDA archives. The ELSST has been integrated with the UKDA dataset database. It is desirable that a similar mapping be achieved between the ELSST and resources deposited into the DAMES repository.

### 3.13 Text Encoding Initiative (TEI)

The Text Encoding Initiative (TEI) represents digitised text documents. It is maintained by an international consortium of academic institutions with a primary focus on documents in the humanities and social science. TEI is more appropriate for qualitative rather than quantitative data, and as such is not appropriate for this stage of DAMES.

# 4 The DAMES Solution and Future Work

The DDI standard version 3.0 has been selected for the DAMES project because it is the most appropriate for the resources exploited in DAMES, supports a migration path for existing GEODE metadata, and ensures compatibility with existing social science practices. In addition, DAMES will examine the use of the ELSST for semantically-based data discovery and processing.

DAMES will provide services for variable operationalisation and dataset linkage. These will involve resource discovery through XQuery metadata queries, automatic generation of metadata during data transformation procedures, and metadata resource descriptions. DDI 3 metadata schemas that group and compare datasets will be investigated as the data linkage descriptor. Work is under way to design an appropriate DDI 3 profile for DAMES resources and to create a "metadata wizard" to help users define metadata. To best support GEODE resources, mappings from its DDI 2-based metadata to DAMES DDI 3-based profile will be developed.

The DAMES e-infrastructure will include procedures for users to register and deposit resources. Resource registration will include a mix of auto-generated and manually secorded metadata description. Deposited and generated metadata will be stored in an XML database. Initial experiments have been conducted using eXist-db (exist.sourceforge.net), and it is likely that this will be used to store the DAMES metadata. Investigations have unveiled challenges pertaining to scalability, robustness and security. The AMGA metadata catalogue (Arda Metadata Grid Application, amga.web.cern.ch/amga) has been used to store metadata in grid environments [24]. Future work on metadata storage is planned to include an examination of providing AMGA services using eXist-db.

# References


[1] Beedham, H. and Missen, J. and Palmer, M. and Ruusalepp, R. Assessment of UKDA and TNA compliance with OAIS and METS standards, Colchester: UK Data Archive, vol. 30, 2005.

[2] Lambert, P. and Tan, L. and Turner, K. and Gayle, V. and Prandy, K. and Sinnott R. Developing a Grid Enabled Occupational Data Environment, Proc. 2nd International Conference on eSocial Science, Manchester, June 2006.

[3] Tan, L. and Gayle, V. and Lambert, P. and Sinnott, R. and Turner, K. GEODE - Sharing Occupational Data Through The Grid, Proc. 5th UK eScience All Hands Meeting, Nottingham, September 2006.

[4] Data Documentation Initiative (DDI), http://www.icpsr.umich.edu/DDI/, Feb. 2006.

[5] Russell, M. and Novotny, J. and Wehrens, O. GridSpheres Grid Portlets. Computational Methods In Science and Technology, 12(1):89-97, 2006.

[6] OGSA-DAI Open Grid Service Architecture, Data Access and Integration, http://www.ogsadai.org.uk, Feb 2006.

[7] Globus Alliance. GT 4.0 Java WS Core : Users Guide. http://www.globus.org/toolkit/docs/4.0/common/javawscore/user-index.html, accessed November 2008.

[8] Object Management Group. http://www.omg.org/, accessed September, 2008.

[9] Object Management Group, Common Warehouse Metamodel Specification, Version 1.1, March, 2003.

[10] Bradley, B. Whither DDI? Situation and Prospects in Canada. Part 2: Making the DDI Work – Assessment and Recommendations, Human Resources and Social Development Canada, 2007.

[11] Thomas, W. and Gregory, A. and Gager, J. and Kuo I.-L. and Wackerow A. and Nelson C. Data Documentation Initiative. Technical Report Version 3.0, DDI Alliance, Michigan, USA, Apr. 2008.

[12] Hillmann D. Using Dublin Core, Dublin Core Metadata Initiative, http://dublincore.org/documents/2005/11/07/usageguide/, 2005.

[13] Library of Congress, Development of the Encoded Archival Description DTD, http://www.loc.gov/ead/eaddev.html, 2002.

[14] Cabinet Office, e-Government Metadata Standard Version 3.1, e-Government Unit, Technical Policy Team, Metadata Policy Co-ordinator, 2006.

[15] Library of Congress, Understanding MARC Bibliographic: Machine-Readable Cataloging, Network Development and MARC Standards Office, http://www.loc.gov/marc/umb/, 2003.

[16] Balkan, L. et al. ELSST: a broad-based Multilingual Thesaurus for the Social Sciences, Third International Conference on language resources and evaluation, pp. 1873-1877, http://gandalf.aksis.uib.no/lrec2002/pdf/3.pdf, 2002.

[17] Library of Congress, Understanding MARC authority records. 2nd ed, Network Development and MARC Standards Office, http://www.loc.gov/marc/uma/, 2004.

[18] Library of Congress, Metadata Encoding and Transmission Standard: Primer and Reference Manual, METS Editorial Board, Digital Library Federation, 2007.

[19] Lagoze, C. and Van de Sompel, H. The Open Archives Initiative: Building a low-barrier interoperability framework, Proc. 1st ACM/IEEE-CS joint conference on Digital libraries, pp. 54–62, ACM Press New York, 2001.

[20] Schindler, U. and Bräuer, B. and Diepenbroek, M. Data Information Service based on Open Archives Initiative Protocols and Apache Lucene, German e-Science Conference, 2007

[21] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1 Blue Book, http://public.ccsds.org/publications/archive/650x0b1.pdf, 2002.

[22] Arofan, G. and Heus, P. DDI and SDMX: Complementary, Not Competing, Standards, Open Data Foundation, http://www.opendatafoundation.org/papers/DDI_and_SDMX.pdf, 2007.

[23] DDI Alliance, DDI-SRG Consideration of other Metadata Standards, Minutes of the DDI structural reform group, http://www.icpsr.umich.edu/DDI/committee-info/minutes/2004-10-24.html, October 2004.

[24] Santos, N. and Koblitz, B. Metadata services on the grid, Proc. of Advanced Computing and Analysis Techniques (ACAT'05), Zeuthen, Berlin, http://amga.web.cern.ch/amga/publications/nsantos05Metadata.pdf, May 2005.