Development of a conceptual Graphical User Interface Framework for the creation of XML metadata for digital archives

Von der Fakultät für Ingenieurwissenschaften,

Abteilung Informatik und Angewandte Kognitionswissenschaft

der Universität Duisburg-Essen

zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften (Dr.-Ing.),

genehmigte Dissertation

von

Dipl.-Ing. Crispen Mugabe

aus

Rusape, Zimbabwe

1. Gutachter: Prof. Dr. Wolfram Luther

2. Gutachter: Prof. i. K. Dr. Andreas Harrer

Tag der mündlichen Prüfung: 20. September 2012

Abstract

This dissertation is motivated by the DFG sponsored Jonas Cohn Archive digitization project at Steinheim-Institut whose aim was to preserve and provide digital access to structured handwritten historical archive material highlighting New Kantian philosophy scattered in the correspondence, diaries and private journals kept by and written to and by Jonas Cohn.

The dissertation describes a framework for processing and presenting multi-standard digital archive material. A set of standard markup schema and semantic bibliographic descriptions have been chosen to illustrate the multiple standard and hence semantic heterogeneous digital archiving process. The standards include Text Encoding Initiative (TEI), Metadata Encoding and Transmission Standard (METS) and Metadata Object Description Schema (MODS). The chosen standards best illustrate the structural contrast between the systematic archive, digitized archive and digitized text standards. Furthermore, combined digital preservation and presentation approaches offer not only the digitized texts but also metadata structured variably sized images of the archive documents enabling virtual visualization. State of the art applications focus solely on either one of the structural areas neglecting the compound idea of a virtual digital archive.

The content of this work describes the requirements analysis for managing multi-structured and therefore multi-standard digital archival artefacts in textual and image form. In addition to the architecture and design, an infrastructure suitable for processing, managing and presenting such scholarly archives is sought for recognition as a digital framework useful for the preservation and access to digitized cultural resources. The proposed solution therefore includes the instrumentation of a conglomerate of existing and novel XML technology for transformations based in a centralized application. The archive can then be managed via a client-server application thereby focusing archival activities on structured data collection and information preservation illustrated in the dissertation process by the:

- Development of a prototype data model allowing the integration of the relevant markup schema
- Implementation of a prototype client server application handling archive processing, management and presentation and based on the data model already mentioned
- Development and implementation of a role archive access user interface

Furthermore as an infrastructural development serving expert archivists from the humanities, the dissertation explores methods of binding the existing XML metadata creation process to other programming languages. In doing so, one opens further for channels simplifying the metadata creation process by integrating the use of graphical user interfaces. To this end the java programming language, its swing and AWT graphical user interface libraries, associated relational persistency and enterprise client server architecture resemble a suitable environment for integrating XML metadata into main stream computing. Hence the implementation of Java XML Data Binding as part of the metadata creation framework is part and parcel of the proposed solution.

Zusammenfassung

Diese Arbeit geht hervor aus dem von der DFG geförderten Projekt zu Digitalisierung des Jonas Cohn Archivs im Steinheim-Institut, dessen Ziel es ist, eine strukturierte Auswahl von Handschriften des Philosophen Jonas Cohns in digitaler Form zu bewahren und den Zugang zu ihnen zu erleichtern.

Die Dissertation beschreibt ein Rahmenwerk für die digitale Verarbeitung und Präsentation digitalisierter Archivinhalte und ihrer Metadaten, strukturiert anhand von mehr als einem Beschreibungsstandard. Eine Auswahl von Standard Markup Schemata und bibliographisch semantischen Beschreibungen wurde getroffen, um die Problematik darzustellen, die aus der Berücksichtigung mehrerer Standards und damit aus semantischer Heterogenität des Digitalisierungsprozesses entsteht. Diese Auswahl umfasst unter anderem die Text Encoding Initiative (TEI), Metadata Encoding and Transmission Schema (METS) und Metadata Object Description Schema (MODS) als Beispiele für Beschreibungsstandards. Diese Standards sind am besten geeignet, die strukturellen und semantischen Unterschiede zwischen den Standards eines systematisch und semantisch zu digitalisierenden Archivs darzustellen. Zusätzlich verbindet der Ansatz die digitale Bewahrung und Präsentation von digitalisierten Texten und von Metadaten strukturierter Bilder der Archivinhalte. Dies ermöglicht eine virtuelle Präsentation des digitalen Archivs. Eine große Zahl bekannter Digitalisierungsanwendungen folgt nur einer der beiden Strukturierungsziele Bewahrung und Präsentation, wodurch der Ansatz eines vollständig virtuellen digitalen Archivs vernachlässigt wird.

Der Schwerpunkt dieser Arbeit ist die Beschreibung einer Managementinfrastruktur für die Erfassung und Auszeichnung von Multi-Standard Metadaten für digitale Handschriftensammlungen. Zusätzlich zu der Architektur und dem Design wird nach einer geeigneten Infrastruktur gesucht für die Erfassung, Verarbeitung und die Präsentation wissenschaftlicher Archive als digitales Rahmenwerk für den Zugang zu und die Bewahrung von Kulturbesitz.

Die hier vorgeschlagene Lösung sieht deshalb die Nutzung bestehender und neuer XML-Technologien vor, verknüpft in einer zentralen Anwendung. So wird im Rahmen der

Dissertation die Strukturierung des Archivs mittels einer Client-Server-Anwendung betrieben und die Bewahrungsmaßnahmen als Prozess herausgearbeitet. Die Arbeit verfolgt mehrere Zielsetzungen:

- Die Entwicklung eines prototypischen Datenmodells mit der Einbindung relevanter Markup Schemata
- Die Implementierung einer prototypischen Client Server Anwendung für die Bearbeitung, Erfassung und Präsentation der Archive anhand des beschriebenen Datenmodells
- Die Entwicklung, Implementierung und Bewertung einer Benutzerschnittstelle für die Interaktion mit dem Rahmenwerk anhand einer Expertenevaluation.

List of Figures

List of Tables

Acknowledgements

I'd like to express my sincere appreciation and thank my supervisors Prof. Luther and Prof. Harrer for their assistance and advice throughout the dissertation process. I consider myself very fortunate to have known and benefited from their academic knowledge and culture. Their supervision is worthy of the German language reference to “Doktorvater”.

The Steinheim-Institut executive and Prof. Michael Brocke I’d like to thank together with Dr. Margret Heitmann the curator of the Jonas Cohn Archive for their commitment and for making my research possible by allowing me to participate in the digitization process of the Jonas Cohn archive. It is through my involvement in this digitization project that I acknowledged and tackled a research problem visible to me and seen worthy of doctoral research. At the same time I wish to express my gratitude to the DFG who funded the digitization project and the University of Duisburg-Essen for the scholarship assistance which enabled me a concentrated environment to finish off the dissertation work.

The evaluations and testing associated with my research work would not have been possible had it not been for a selection of colleagues who invested their time and resources allowing me access to their working methods, archives and artifacts. To this end I’d like to thank Dr. Jobst Paul of Duisburger Institut für Sprach und Sozialforschung (DISS), PD Dr. Wolfgang Treue of the Alliance Israélite Universelle project, Dr. Britta Caspers and Dr. Christoph Bauer of the Hegel Archiv Bochum in addition to Nathanja Hüttenmeister, Thomas Kollatz, Karina Küser and Ann-Kathrin Heidenreich from Steinheim-Institut who all helped as evaluation test users and in the evaluation process. The colleagues from Steinheim-Institut and beyond I’d like to thank for their support be it logistical or simple encouragement and their motivating words.

For my participation in a selection of archival and digitization activities I’d like to thank the Allegro HANS group, Kalliope and the Staatsbibliothek zu Berlin, Dr. Eva Dyllong and the Retrodigitalisierung Workshop team and finally the InterFace2011 committee of the University College of London and their editorial colleagues of the Oxford Journal for Literary and Computing Linguistics. My participation and the access to the scholarly and academic interaction activities they organized gave me an insight into the current state-of-the-art and research activities in the digital humanities and information sciences field.

Lastly and most importantly I’d wish to extend my sincere thanks and gratitude to my family, my daughter Clara and my wife Petra as they sacrificed a great deal in order for

me to realize my dream of doing doctoral research and writing up a dissertation. As such I dedicate the dissertation to Clara as it has been part and parcel of her family life accompanying her day to day since birth.

Table of Content

1 Introduction

The proliferation of information via digital systems has since the discovery of hypertext become “the” characteristic of the modern electronic age catapulting electronic media to being a part of everyday life. However, one appreciates the traditional backbone of information transmission and storage as being written text. In the same manner in which written text has developed from rock paintings, hieroglyphics to modern day writings so has the media via which the information is stored and accessed commencing from rocks, papyrus, paper and now digital computers. Acknowledging this development reveals the natural consequence of not only the preservation but also accessibility of historical texts to a modern information society

Traditionally archives and libraries have taken up the role of being information silos preserving information i.e. stored texts. In the same respect catalogues have played a major part in structuring these information silos and facilitating access to the stored information. Likewise, catalogues are written information sources using the same information storage and proliferation resources as the information stored in the library and therefore also exposed to the respective changes in information and communication technologies. The result for the digital information age being that archive management is transformed from the card index catalogue to the digital library.

In addition to developments in digital text and information processing, the spread and popularity of the internet and the introduction of the worldwide web service opened new and novel information dissemination prospects, one of these prospects being the virtual digital library. Not only can the libraries and archives serve a global audience with catalogue information but also with virtual objects i.e. original books, documents or artifacts whose access was earlier preserved for a chosen few, in addition to further services such as watermarked electronic copies, document transcription and translation.

In light of these prospects and a worldwide audience there is need to develop management systems for virtual information silos enabling information access, management and interchange. This work presents a graphical interface and a framework which simplifies this process whilst integrating the structural formats for catalogues, text and the respective image objects and data export.

1.1 Digitization

The transformation of written or printed information into structured and coded machine readable information can be described as digitization. Digital data is then “a sampling of original data encoded” [CR06] understood by a computer and liable to automated processing. Applying this to digital libraries, whose definition refers to a collection of digital objects, extends the definition to include the process of transforming real library or archive objects into virtual digital objects by character recognition e.g. with OCR, transcription and facsimile editions. The digital library becoming digital information organized in a database or marked-up in structured languages [CR06]. Given the digital age in which we are now, historical artefacts transformed into their digital form as images with associated text transcriptions can now be digitally preserved for current and future generations in electronic repositories defining the process of digitization in electronic collections and archives. Given this scenario the structuring of the electronic archive using standard metadata becomes imperative. The standards should address:

- Descriptions within a digital library e.g. with METS
- Collection-level descriptions e.g. search aids with Encoded Archival Description (EAD)
- Transcribed electronic texts e.g. with Text Encoding Initiative (TEI)

Metadata

Describing the different aspects to be understood by the term digitization leads to a further problem to be addressed in this dissertation i.e. metadata exchange. As a result of the different tasks associated with a particular standard or the upgrading to modern standards metadata have to be either extracted from a digital library and restructured or mapped directly to another standard. The associated methods include schema based transformations for XML conform metadata, repository oriented solutions, web-crawling and metadata crosswalks for parallel categories. The selected methods depend on the metadata tasks e.g. it would be difficult to crosswalk electronic text metadata such as in TEI because these texts are more tagged data as opposed to catalogued description data. Frameworks for description metadata will can be implemented and transformations into different schemes via XSLT possible.

Presentation

The problem associated with presenting digital information on the internet often centre on awareness, digital archives are no exception. The main aim of the presentation is therefore to develop user interfaces that resemble real archives or library structures, where possible answering the question of provenance [CR06] using metadata. The classical starting point is with the digital objects themselves. The presentation is oriented with the document type i.e. hierarchically structured, plain unstructured text documents, page images (with and without extracted text) or metadata [WB03]. On the other hand resembling a real archive implies different modes of accessing the digital information. The interfaces need to be modeled according to user roles and classical model-driven approaches for multiple user interfaces could play a role in achieving this goal [BO06].

1.2 Structured Encoding

From the preceding sections of this chapter it can be derived that digital data processing plays a pivotal role in this dissertation. In general, data can be seen as a compound element referring to *"encoded stimuli that convey meaning"* [EM79] [SD79]. Depending on the application area, and the semiotic classification of the data, it can then be declassified according to purpose into the subcategories message, signal or information. The diagram in Fig.1.1 below adapted from Dworatschek's semiotic code analysis [SD79] classifies an information system at an abstraction level on the basis of the suitability of a selection symbols to convey information. The defined abstraction levels semantics, syntactic and pragmatics qualify the chain of symbols as being either information, a message or simply a signal.

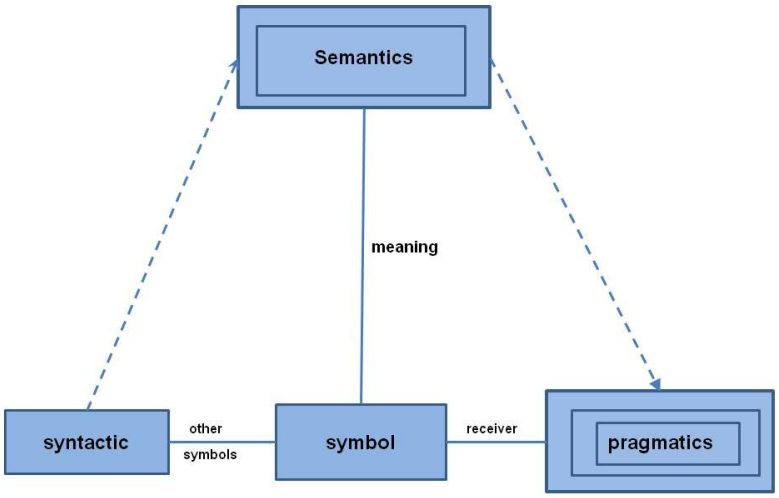


Fig. 1.1: Semiotic classification of a system adapted from [SD79])

Information <=> Pragmatic

Message <=> Semantic

Signal <=> Syntactic

The resulting ambiguity between the terms data processing and information processing can be resolved by specifying the concept of information and its relation to the compound notion of data. In this case, the principle of *information* as to be understood in this work bundles the semantic and syntactic semiotic sub concepts into the semiotic pragmatic form relating the structure, content and impact on the participants of an interacting environment i.e. a system [SD79]. The compound relation between elements of the interacting environment is then communication which in turn takes place on the backbone of *encoded* signals (i.e. syntactic data). Hence the notion of *encoding data* extends the principle of *information* with reference to its dissemination and role in an environment of interacting objects. In which case information can be seen in a *systems frame of reference* as *serving either of two roles or functions 8* [EM79] i.e. either as *system input* contributing towards an output or as the *consolidated system output* of some transmitted or processed input [EM79]. A refinement of *information* by means of *encoding* stimulates an understanding of the system inputs and outputs by the interacting environment facilitating *consolidation*.

In other words, encoding ensures that the receiver understands that what the sender has transmitted to him and hence represents a mutual language between the interacting objects i.e. the sender and the receiver and a description of the system data resources. Subsequently, the processing of information in a system by organizing the mutual language with the output further serving as system input and the mutual language representing the set of relations between the interacting objects can be referred to as structured encoding or resource description.

Structured Markup

Archive digitization activities resemble the transfer of collection activities from paper based preservation via isolated machine-readable systems to modern state-of-the-art client server web-based systems. As such, digital archive resources and information resemble electronic documents whose syntactic information can be highlighted by a generalized markup language understood by clients and servers interacting within a system, a summary of which is illustrated below.

Standardized General Markup Language SGML

Derived from the Generalized Markup Language GML and an ISO 8879 information processing standard, SGML is a meta-language designed with shared long-term preservable machine readable electronic documents in mind [W3C]. As the name suggests SGML is a general language mutual to text based systems on a syntactical level and describing the structural composition of text based documents in preparation for further use as *information input* for text resource processing and sharing systems. It is therefore no surprise that emphasis is placed on validity, whilst aspects of *information meaning* are left to semantic level languages in the prologue e.g. DTD or DSSSL (Document Style Semantics and Specification Language).

Hypertext Markup Language HTML

Originally formulated as an SGML *"application"* denoting the structural semantics of content in text and in so doing facilitating the creation of structured documents, HTML represents the predominant mutual markup language understood by web browsers. Although it also highlights meta-information, mainly the semiotic pragmatic system input data described by HTML is rendered content and illustrated by the receiving web browser.

Extensible Markup Language XML

Defined as a meta-language and developed as a text-based derivate of SGML representing structured information and suitable for web use [W3C], XML has proved itself to be an excellent and widely accepted medium for information interchange and representation.

The formal definition of XML itself ranges from being “the description of a class of data objects, so called XML-Documents” [KST02], to a “very flexible text format derived from SGML designed to meet the challenges of large scale publishing” [W3C]. The definitions more or less describe the role XML plays in the accumulation, preservation, presentation and interchange of large amounts of data. The description of these roles can be mapped to the roles of an archive, museum, library or any institution associated with the collection, preservation and presentation of material. The common aspects though, include the notion of a common encoding language (set of rules) and machine-readability of the documents encoded using this common language. The reference to *information sharing* illustrates XML semiotic pragmatism justifying classification of XML as a markup language.

Extensible Hypertext Markup Language XHTML

As an extended version of HTML formulated as an XML *“application”* and serving as meta-language for the structuring and semantic markup XHTML is considered to belong to the family of XML markup languages. With a focus on interoperability and extensibility, XHTML reflects its derivation in XML syntax, the requirement of well formed structures and modularization for sub-setting and extension. XHTML encoding also resembles pragmatic input for semantic web systems of which the extended XHTML + RDFa assumes the role of the semantic meta-language.

Resource Description Framework RDF

Officially, RDF is defined by the World Wide Web Consortium as a *“standard model for data interchange on the web”* [W3C] i.e. a metadata data model. In practice however, it has assumed the role of a model for the formal description of web information resources identified as objects predestined as implicit machine-readable input for further processing applications. Characteristic is the relational linking structures referred to as *“triple”* and the related formal semantics based on labeled directed graphs. In other words Resource

Description Framework (RDF) resembles the foundations of structural description languages e.g. Ontology Web Language OWL and Simple Knowledge Organisation System SKOS; upon which web ontologies can be defined preliminary to data integration and interoperability activities descriptive communities" [W3C].

SGML	*METS*
<entry> *<hwsec>* *<hwiem>bungler</hwiem>* *<pron>b<l> g</>ngler</pron>. </hwgp>* *<vfl>Also<vd>g</vd> <vf>bungler</vf>,* *</vfl>* *<etym>f. as prec. + <xra><xlem> -ER</xlem>* *<sen>One who bungles; a clumsy unskillful <quot>* *<qdat>1533</qdat>* *<auth>MORE </auth>*	*<mets:structMap TYPE="PHYSICAL">* *<METS.DIV id="PHYS92081" DMD ID="md92081"* *ADMID="amd92081"<mets:divID="div931-T-I-01" ORDER="1">* *<mets:fptr FILEID="img946-RT-1-01"></mets:fptr>* *</mets:div><mets:div ID="div931-T-I-02" ORDER="2">* *<mets:fptr FILEID="img946-RT-1-01"></mets:fptr>* *</mets:div>*

Fig 1.2 SGML metadata vs. METS structural metadata

Metadata Encoding and Transmission Standard METS

METS is an XML Schema designed as a standard digital library metadata encoding language describing administrative, descriptive and structural metadata. METS focuses on digital objects and their hierarchical structure embedding other metadata formats e.g. MARC, MODS whilst associating the structural maps with object instances and file locations. The seven characteristic sections of a METS document enable the modeling of real objects e.g. books, collections, manuscripts or bibliographic records utilizing the located digitized objects and the structural hierarchy as resembled in the real objects. Fig. 1.2 above shows SGML and XML fragments illustrating structured markup using the aforementioned markup languages. In this particular case the element tags of the respective markup languages host the machine readable syntactic information. The extensible character of the XML language is highlighted by the syntactic tags <mets>

further structuring the data in the METS fragment in accordance with the METS standard. In addition to bringing in useful data processing advantages, the structuring as seen in a multi-standard scenario may complicate the encoding process and is core to the digital archiving challenges addressed by the dissertation.

Problem Description

Digital Archives

A novel mode of preserving, presenting and accessing cultural heritage is through the construction of digital libraries and archives. These organized collections of information [WB03] provide an electronic platform for exchanging structured data and exploring the semantics of the data in question. Witten et al.'s definition of a digital library refers to “a focused collection of digital objects...along with methods for access, retrieval and for selection, organization and maintenance of the collection” [WB03] clearly outlining characteristics which qualify a digital library. The effectiveness and acceptance of such a digital library is based on other criteria namely users, objects and technology as summarized in Bishop et al.'s definition “A successful digital library is a place where a group of users *(people)* can effectively search a group of documents *(collection)* via an information system *(technology)*. These three components must be in harmony” [BV03]. The end users can be distributed into subsets:

- **Administrators** who maintain, organize and administer collections and
- **Users** who access and retrieve the contents of the library

The contents in question are documents, also referred to as digital information objects and metadata which describe the documents. Metadata is considered to amplify bibliographic cataloguing practices in electronic environments and can be can be classified in one or more of the following functional categories [OR01]:

- **Descriptive:** facilitating resource discovery and identification
- **Administrative:** supporting resource management within a collection
- **Structural:** binding together the components of complex information objects

Digital Libraries have, to date, mainly focused on and applied descriptive and structural metadata as a result of the successful Dublin Core Initiative [DC09] and a multitude of other descriptive metadata standards ranging from Encoded Archival Description [EAD09] to Text Encoding Initiative [TEI09]. Hence, the administrator or librarian structures information objects in the library using metadata, thereby assisting the user in his quest to search and retrieve documents in the library. Given the nature of preservation, a

preservation digital library is required to provide copies of its content to federal archives resulting in object and metadata exchange between libraries.

Cataloguing

Traditionally libraries and archives digital or not serve to be knowledge silos storing information for the future from a time discrete point of view. In order to serve this purpose well, knowledge management techniques have been in use for ages serving as aids for the quick access and interpretation of the knowledge at hand. The term knowledge is itself often subject to debate in relation to its differentiation from the terms data and information nevertheless, importance still lies in its structuring and dissemination and clarification is obtained after the former have been implemented. Traditionally categories play the biggest role in structured information sources and with digitization this process of categorization and structuring for computer based processing and presentation is now described as knowledge engineering. It is therefore no surprise that the multitude of digitization projects within the notion of digital libraries and archives is mainly involved with categorization using standard library tools such as allegro, greenstone or archivists toolkit and the often reference to this practice as digitization. The digitization implies the transfer of catalogue information from catalogue cards to digital machine-readable (online-) catalogues. There are some abstract windowed machine-readable metadata capturing tools encoding metadata in the classical library standard MARC, which has been updated to include MARCXML. However, they do not include a cross walking facility or any interface to facilitate interoperability and participation in open initiatives is more of a byproduct. Defining an interface and framework for abstract cataloguing as part of an integrated digital infrastructure is part of the problem to be addressed by my dissertation.

Digitization as Preservation

The term digitization is often associated with preservation of endangered archive resources by either photographic means or a facsimile of the archival objects in question. For most paper based archival objects the poised danger results from oxidation and exposure to light in addition to the decomposition of the paper material. Therefore as long as the objects are still in a relatively good state, preservation concentrates on reducing or in effect a complete denial of access to the original document whilst maintaining access to its contents, now available in digital photographic form. This phenomenon is not restricted

to historic archives, business document management systems process and archive business correspondence and objects in the same manner.

Character Recognition

The most common form of digitization is in the form of optical character recognition OCR; here pattern recognition filters are implemented on scanned images of texts i.e. handwritten, typed or printed producing machine-encoded and hence readable text as a result. In practice commercial OCR tools either standalone or integrated in scanners are widely available however, OCR has proven to be restricted since it requires the calibration of each font to be recognized. Whereas this requirement proves to be trivial for typed or printed text it requires an enormous effort when applied to handwritten or calligraphic texts. This phenomenon also applies to the Jonas Cohn Archive and its contents handwritten in Jonas Cohn's *sütterlin* type handwriting. Not only the calibration of Jonas Cohn's handwriting but also the calibration of its variations associated with the author's age, environment and state of health require enormous effort and a budget beyond that of the digitization project as a whole. Traditionally, digitization of such handwritten texts is assisted by human manpower in the form of transcription resulting in typed fonts either machine-encoded or as a forerunner for OCR based processing.

Transcription

In addition to bibliographic records and facsimile images, transcription also represents digitization particularly when dealing with non-standardized text e.g. handwritten or calligraphies. Transcribed text archives serve as the basis for further encoding in accordance with guidelines such as TEI, EAD etc. Transcription environments can be embedded into a digitization framework in addition to the XML document instances which as a rule denote digital library object structure [TEI09].

Integrated Digital Archives

An integrated approach as proposed and described in this dissertation views the digital archive as an integrated entity contain bibliographic records, text and images. The integrated digital archive then serves the combined purposes of digital preservation, digital record keeping and digital archive presentation and access. Now each of the purposes mentioned above is associated to and in certain cases bound to a particular metadata

structure associated with a specific metadata schema and at the same time encodes common information about the same archive to a different audience. Therefore, the task of any framework assisting the creation of the metadata includes the "*predefining of the overall structure and capturing common design decisions"* [EG95] i.e. an abstraction of archive metadata structure from the encoding.
Gamma et al. [EG95] speak of *"emphasize design reuse over* code reuse*"* and with technology rapid technology changes this forms the basis for future machine-based processing.

Summary

Given the electronic age in which we live in, it is imperative that archive material be electronically collected, thereby easing preservation and access, both for current and future generation, data structuring for easier search and data interchange and presentation making use of modern presentation media such as the internet. In light of this, the aim of this dissertation is to develop an application for collecting and managing common digital resources of archives, drawing capabilities from the well of XML-based tools to represent and exchange archive collection data. Given that archivists, librarians and curators are generally not computer scientist, this application should attend to their needs by providing a graphic user interface enabling them to tag, link and transform their data into and using the relevant XML standard formats.

1.3 Scope and Goals

In this dissertation, a conceptual graphical user interface assisted framework for structuring and encoding heterogeneous metadata for digital archives is introduced. The main feature of the conceptual framework is an abstract encoding of heterogeneous digital objects (text, facsimile, records) metadata in a uniform structure and a consequent crosswalking into relevant XML schema. The graphical user interface serves abstract windowing of the encoding process of which the latter represents the structuring of the digital collection and its metadata and is characterized by specified XML tag sets. The extensible framework has room to accommodate transcription objects and structural mapping of archive objects without necessarily duplicating the object metadata.

Scope

The dissertation is a contribution to semantic digitization in general and *digital archive engineering* in particular. It illustrates how concepts of abstract frameworks implemented in digital archives, can separate archive structure and encoding and in so doing enhance structural compatibility and schema independence. Subsequently, issues such as usability and interoperability derive profit from this abstraction and hence promote structural encoding activities among archivists. Further web engineering issues regarding pattern reusable designs [EG95] and evaluation are also subsequent. The motivation behind this dissertation lies within the project towards Retrodigitization of the Jonas Cohn Archive at the Salomon Ludwig Steinheim-Institut. The goal of that project was a structured digital archive of the handwritten manuscripts for preservation purposes as well as in preparation (crossmedia publishing) for a book edition. As a result heterogeneous object sets i.e. facsimile, manuscript records and text. The scope of this dissertation is to develop a conceptual set of classes specifying the object oriented structure of the digital archive with an exemplary schema mapping.

Goals

The main goal of this dissertation is to contribute towards the simplification and encouragement of the process of creating structured metadata for integrated digital archives with a focus on interoperability and data interchange.

In other words, the development of systems which encourage a guided encoding process fulfilling the standard digitization goals i.e. machine readability and processing of archive contents, improved accessibility and preservation as a result of the former without the user necessarily having to be an encoding expert but at the same time maintaining encoding standards and an XML interface for exporting the structured archive content in a chosen standardised schema i.e. mapping or crosswalking the content to e.g. METS, TEI, MARCXML. Subsequently, the use of standardised mark-up and web technology for structuring and publishing archive content and thereby preparing the ground work for future archive use be it a further processing of the structured texts for semantic purposes or for publication using another media i.e. cross-media publishing. Furthermore, the implementation of pattern and design reuse principles and an introduction of object orientation principles into the encoding process becomes a sub goal of the dissertation in

line with the abstraction principles associated with the computer-human interface and the associated contribution towards framework development, schema compatibility and the avoidance of encoding repetitions of common elements.

Contribution

Contrary to existing frameworks, this dissertation does not aim to develop or elaborate on a descriptive vocabulary or format for digital archives. Instead focus is on the novel notion of creating heterogeneous metadata for interdisciplinary archive description.

The heterogeneity incorporates object and text encoding with record collection and archival description hence the implementation of interdisciplinary metadata encoding standards and XML-Formats. The Jonas Cohn Archive as a case study illustrates this heterogeneity. Furthermore, the centralisation of encoding, interoperability and inventory registration activities by the framework limit the duplication of tasks. Existing frameworks e.g. Tustep, Allegro HANS or TEI's Roma tend to focus on metadata in the isolated homogeneous context of text, object or record encoding and within the restricted disciplines computer philology; library and preservation sciences or web technology.

1.4 Dissertation Outline

The structural organisation of this dissertation consists of three sections, Part I, II and III; an outline of which is summarized in Fig 1.3 The outline is subsequent to the introductory chapter 1 and therefore commences on Chapter 2.

Part I is dedicated to an analysis of postulated problem limiting the scope of the relevant metadata, structured mark-up elements and schema; and their state of the art. The result is a specification of metadata requirements for digital archives and the formulation of specified requirements for the metadata creation framework and the respective mapping and crosswalking facilities. An analysis of the requirements of heterogeneous metadata for digital archives precedes the specification providing orientation, mark-up boundaries and the foundations of a sound data model.

Part II tackles the postulated problem discussing the solution, it's design and subsequently its implementation. The methodical solution in the form of a system model

and the framework classes are to dictate the system architecture and together be implemented as the formulated metadata creation process using the proposed framework and its conceptual user interface on a prototype.

Part III relates to an empirical evaluation of the proposed framework analysing suitability of the framework to digitization activities and usability in general, in addition to references to related work and an outlook on further research and as such summing up to a conclusion of the dissertation.

Part I: Problem and Requirements Analysis

Chapter 2 State of the Art

An analysis and description of metadata types relevant to digital collections in general and this dissertation in particular and their classification according to task commences this chapter. In addition to an assessment of the principles of structured data in general, metadata and their interchange and interoperability in particular, an overview of state of the art structured mark-up and existing XML metadata schema will be described. Furthermore, emphasis on the need for metadata creation and creation frameworks in pre-eminence to interoperability and structured retro-digitization will be made before the need for graphical user interface for the metadata creation framework is introduced. Existing solutions for creating homogeneous and heterogeneous metadata alike including their capabilities and limitations will be outlined. The final analysis summarizes heterogeneous metadata creation framework as an integral constituent of the digitization process.

Chapter 3 System Requirements Analysis

In this chapter focus is on an analysis of the requirements of a digital archive and therefore the requirements of a metadata creation framework. These requirements constitute the backbone of the proposed conceptual framework and resemble the basis upon which normalized model data and process models of the framework are identified. A further refinement of the normalized models identifies the systematic functions specific to metadata creation which in turn relate to the user interface requirements to the metadata creation process specific to digitization of heterogeneous archive collections. All in all, an analysis of the system

requirements as described in this chapter connects the metadata and the creation framework requirements to the task related requirements to be considered whilst modeling system users and their access rights. In other words, the task, data and process models are encapsulated into an integrated system model.

Part II: Solution Design and Implementation

Chapter 4 System Design and Architecture

A description of the system design and the conceptual framework based system architecture are among the key aspects dealt with in this chapter. The metadata creation infrastructure for digital archives comprised of the conceptual framework and the user interface are presented. In addition to the fundamental reasoning and methodology behind the presented design, the chapter illustrates how the developed conceptual framework fulfills as a solution, the specified metadata creation requirements identified in chapter 3. Furthermore, a description of both the architecture enabling the implementation of framework and the implementation as a prototype are presented. The role of reusable and design pattern based software in relation to encoding abstraction and multi-schema metadata creation in the infrastructure to be developed is also described.

Chapter 5 Archive Use Cases

This final chapter in the development stage is dedicated to the description of the use cases in preparation implementation of the design and the evaluation of the metadata creation infrastructure. The description also looks at the existing metadata structures of the use cases thereby also analyzing their user interface requirements to be supported by the metadata creation framework. Further aspects include structural support as well as existing methodology and frameworks for metadata interoperability and interchange. Feasibility and usefulness of the proposed conceptual framework also play a major role, the foundations for which lie in this chapter. The subsequent *proof of concept* prototype is then implemented within the framework of a formative evaluation whose result contributes towards the final implementation.

Part III: Evaluation and Conclusion

Chapter 6 Evaluation

This chapter describes the evaluation approaches and the evaluation framework applied to the graphical user interface supported metadata creating system. Assessment aspects cover both the *learning* process of assessing the target users' needs and incorporating them into the software development process, as well as an empirical verification of the *fulfillment* of the outlined goal of supporting structured digitization and multi-schema data exports. The evaluation focuses on the adequacy of the developed framework in addition to an assessment of the overall usability and attractiveness of the framework and its graphical user interface to archivists and users of integrated digital archives.

Chapter 7 Conclusion and Outlook

A summary of the results of this research work and recommendations for further research are outlined in this chapter. The summary includes the main response to the main research question and goal and the implications for future digitization projects. Figure 1.3 below summarizes the dissertation outline.

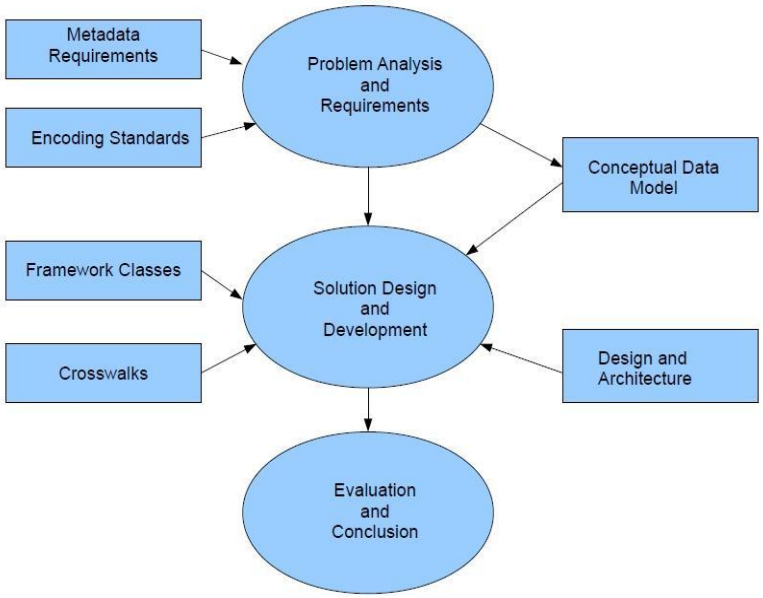


Fig. 1.3 Dissertation Outline

2 State of the Art

An analysis of the most recent ideas, methods and techniques is prerequisite prior to any software development process and this chapter addresses this prerequisite for archive metadata creation systems. To begin with, we will look at the “state of the art” information organization and structuring for digital libraries in general and digital archives in particular in section 2.1. The clarification of the terminology and bibliographic ontology will give us an insight to the scope and novelty of the metadata creation framework. General standards for digital archive metadata and their types follow in section 2.2 in addition to interoperability and metadata schemes for digital library and archive content in 2.3. This dissertation does not however, aim to develop a new encoding scheme for digital archives, rather it seeks to unify and simplify the encoding process based on prevailing encoding schemes. “A focus of research is thus to provide integrated methodologies and tools for presenting and managing digitized archive holdings without them losing context” [DA10]. A discussion analyzing these tools and methods should give us insight as to the extent of their support the intended goal of having graphical interfaces simplifying standardized metadata creation processes.

2.1 Metadata Creation Frameworks

The notion of Metadata Creation Frameworks which is of interest to us, is that which refers to structures supporting and simplifying the process of providing meaning and structure to digitized archive content and data thereby allowing data to be shared and reused across application and collection boundaries. As a result these structures reflect an integration of the concepts of context articulation i.e. resource descriptions and bibliographic organization [DC09] [BV03]. The importance of such frameworks is in line with the theory of user-created metadata alignment to the description of a collection’s elements and its use [BV03], an important factor when digitizing collections whilst focusing on improving their use. This chapter will therefore discuss elements currently used for cataloguing, describing and structuring archive objects and context, subsequently looking at graphical user interfaces supporting such frameworks.

2.1.1 Metadata and Structural Markup

The prefix *"meta"* can in its philosophical context be derived from the concept of "the understanding of knowledge" or linguistically as a set of symbols or language used when describing structure [C11]. Considering that modern day knowledge is stored and distributed via electronic media one comes to the conclusion that the concept of metadata has the role of enabling one to understand knowledge on one hand and on the other hand it represents the language used to describe the structure of knowledge in these electronic systems. In other words metadata in information systems is "data describing data" [DC09] as well as *"machine understandable information for the web"* [W3C]. Gilliland-Swetland et al. describe metadata as a "ubiquitous" term understood differently by the respective *"professional communities that design, create, describe and preserve and use information systems and resources."* serving however, the same goal i.e. the *"development of effective, authoritative, interoperable, scalable and preservable cultural heritage systems"* [GS00]. This implies that depending on the information system in question the role of metadata can be further subdivided among others into resource description [W3C] or summary information on documents in a digital collection.

The former being collected more or less for facilitating access to large information collections [WB03]. In general, metadata are *"data about data"* through which the three basic features of an information object i.e. content, context and structure may be reflected [GS00].

- *Content* intrinsic to an information object and relating to the object's contents
- *Context* extrinsic to an information object relating associations to the object's creation
- *Structure* can be both intrinsic or extrinsic and relating to formal sets of associations

In cultural heritage projects metadata assume the role of structural description aiming at machine understandable language, focused on automated data processing as well as a better understanding of "knowledge". This knowledge which one can describe as structured knowledge remains the key to a successful online cultural heritage presentation

as it influences user models and therefore usability. Borgman [BV03] classifies it as usability based knowledge in the following categories:

- *Conceptual Knowledge* refers to an understanding of the type of information system being used
- *Semantic Knowledge* refers to an understanding of the available steps required to carry out a task
- *Syntactic Knowledge* refers to “an understanding of the commands or actions in a specific system“.

Further aspects of knowledge and its relation to usability will be discussed in chapter 6 on evaluation. However, the relation between collaborative knowledge construction and its three characteristics of being, situated, distributed and social illustrate the inclination of metadata and knowledge creation towards text markup and its role of specifying the structure of documents and controlling their presentation. Cohen et al. [CR06] describe text markup on the basis of its machine readability and the involvement of classification according to the criteria: format, logical structure and context. The same criteria and extensions to include the structural and descriptive aspects are familiar from the metadata definitions described in sections 1.2 and 2.1.1. Although document markup preceded the internet, the use of Meta tags in HTML underlines the inclination mentioned above. In addition to that HTML is a derivative of Standardized Generalized Markup Language SGML, a meta-language and a product of standardization of computerized typesetting, the modern version of the historical manuscript markup for typesetters [WB03] [CR06]. Having determined the relationship between metadata and text markup a further analysis of the markup languages brings us closer to modern metadata creation frameworks. Of interest is the difference between specified languages and meta-languages within text markup activities. Whilst specified languages such as HMTL and XHTML, the underlying document formats for the worldwide web, are designed to allow hyperlinks to other files and serve to structure the presentation of electronic documents. Meta-languages such as SGML and XML serve as the framework for describing document structure and metadata i.e. they are languages which describe other languages and markup formats. Witten et al. al. go further to describe HTML and XHTML as mark-up languages as opposed the meta-

languages SGML and XML with markup identifying metadata in electronic documents and controlling their structure and appearance [WB03]. It is interesting to note though that the tag content section of several meta-languages is syntactically identical to that of HTML.

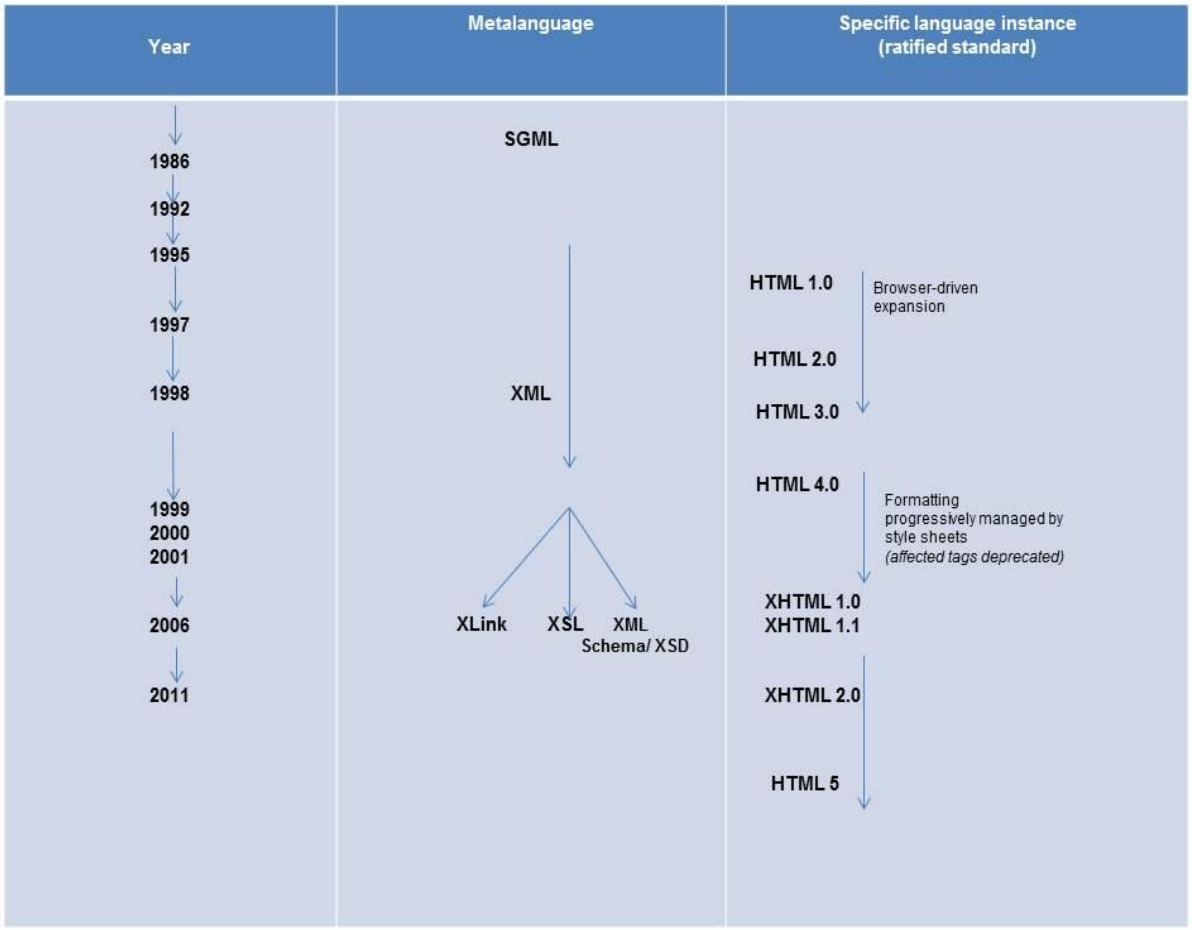


Fig. 2.1 Relationship between SGML, XML, HTML [WB03]

2.1.2 Bibliographic Metadata and Markup

The goal of any digitization project, in addition to the preservation of historical artifacts is the providing quick and useful access to the digitized objects be it via a catalog or a digital presentation both of which require authoring tools. On the other hand creation of library metadata aims to provide both physical and intellectual access to content whereby their creation is subject to cataloging rules as well as structural and content standards [GS00],

the complexity of which can be illustrated by two standard metadata formats i.e. machine-readable cataloging (MARC) and Dublin Core. The former being a development for professionals, is comprehensive and well developed, whilst the latter a development for non and professionals alike, hence minimalist [WB03]. The structure of Bibliographic Metadata originates from bibliographic organization of library systems and is as such oriented on these and their objectives. It is therefore inevitable that metadata creation frameworks are directly associated with catalogue information management.

Bibliographic Objectives

These are intended to reflect user needs and can be evaluated with respect to their sufficiency and necessity. Online bibliographic systems are contrary to traditional indexes and catalogues dealing with online documents with varied degrees of control. The main aim is to develop a system for organizing information in accordance with the five basic bibliographic objectives.

These objectives constitute the hypostatization of user needs representing the transition from catalogues as inventories and finding aids to navigation, structuring and digital preservation tools and are described as follows:

- **location**:
  - finding objective:
    Specifies the location of a particular document
  - collocating objective:
    specification of the location of a set of documents defined by criteria such as author, work or subject.
- **Identification**:
  Distinguish between two or more sought entities with similar characteristics
- **selection**:
  Choose an entity in accordance with user requirements and needs respective to content, physical format etc.
- **acquisition**:

Access entity, in the case of online archives, electronically

- **navigation**:
  Browsing through a set of entities conforming to one or more of the user requirements

Operational Objectives

These are intended to act upon the bibliographic objectives and facilitate interaction with the user and according to the user's needs. Evaluation of the operational objectives is in the framework of usability evaluation with respect to fulfillment of bibliographic standards and user expectations. As a result most online archiving systems are traditionally modeled according to user tasks and the resulting patterns which will be described in section 3 on the system model. The main aim is to identify aspects relevant with regards to the organizing of information in accordance with the bibliographic objectives and user interaction hence summarized as follows.

- entity specification
- attributes specification
- specify relationships

Bibliographic Ontology

In addition to having defined bibliographic objectives and their operational objectives, it is only natural that one defines and describes the bibliographic objects in question and their relation to one another. This process can be described as ontology, *"the study of being"* and a representation of knowledge in a domain as sets of concepts and their relation to one another. Svenonius [S00] extends this definition of ontology to include *theories regarding abstract entities being admitted into a description language*. Given the nature of digitized archives, the main aim of the dissertation can be described summarization of explicit specifications for shared *(concepts)* content and is hence in line with the concept of bibliographic ontology.

Given that a bibliographic or archive theory is characterized by entities which in turn serve as variables for the ontology's scientific theory, these entities summarized in table 2.1, make up the primary objects and abstract admitted into a description language [S00].

Bibliographic Entities	Archival Entities	Digital Objects	Other Entities
• documents • works • editions • authors • titles • subjects	• correspondence • addressee • journal • lecture • sermon • person • addressee	• text • image	• impression • imprint • archive • collection

Table 2.1 Foundations of Descriptive Entities

The role of the concept of ontology in digitized archives is almost natural given the aims and motivations for digitizing archive material. Standardization, centralized archiving & multiple record keeping, information exchange and format crosswalking lead to the question *"when do two descriptions describe the same entity"* [S00] and hence the ontological question what are the objects of a bibliographic (archival) description. In chapter three of this dissertation a model outlining archive entities, attributes and relationships will be built as an approach to ontological question providing a structural framework assisted by a GUI for creating archive metadata.

Description Languages

Understanding the state of the art archive metadata infrastructure requires an understanding of the concept of description languages. This concept defines the need to define information to be organized and has generally been dealt with under the umbrella term "bibliographic languages" [S00] upon which descriptions are recorded on bibliographic records i.e. cards with the resulting description being "a statement of characteristics or relations serving to identify an object". Traditionally a bibliographic description language is classified either as a work language or a document language both of which respectively describe the following attributes i.e. sub-languages:

Work Language

- author language
- title language
- edition language
- subject language
  - classification language
  - index language

Document Language

- production language
- carrier language
- location language

The components of a bibliographic language are then summarized as follows [S00]:

- **vocabulary**: constitutes of a list of terms classified as *derived terms* i.e. taken as is or *assigned terms* i.e. normalizations. The former are descriptive metadata whereas the latter are organizing metadata [S00] and will be elaborated in section 2.2 on metadata types. Vocabulary classification and terminology is according to the bibliographic language in question in line with the following categories:

  - subject language:
    - keywords
    - descriptors
    - index terms
  - work language:
    - data elements
    - metadata

- **bibliographic semantics**: require standardization and via vocabulary control and normalization and provide meaning to structures within the bibliographic language classified according to the following subcategories:

  - relational semantics

  - referential semantics
  - category semantics
- **syntax:** *"the grammatical arrangement of words in a sentence"* [CO11]
- **pragmatics**

Machine-readable Vocabularies

The electronic era and the proliferation of digital libraries and digitized archives have resulted in the need to adapt the bibliographic concepts mentioned above to suit the electronic and internet based environment. The Machine Readable Cataloging (MARC) format ushered in the electronic era moving away from cards to coded metadata elements however in so doing losing the bibliographic structuring of the elements in addition to the *"syndetic"* structures guiding users towards information organization language i.e. navigation guide [S00]. On the other hand developments in markup languages such as XML and the resulting tailor made tagging schemes provide an opportunity to integrate archival content and bibliographic metadata structuring using modern client server and graphical interface technology. This represents the goal of the system developed within the framework of this dissertation in addition to simplifying the creation of heterogeneous metadata irrespective of encoding knowledge. Structuring metadata using markup languages is generally described as bibliographic markup as elaborated below.

Internet Resource Metadata

In addition to the consensus on the role of metadata in library resource discovery is also the acceptance that their efficacy for web based cataloging is insufficient. Gill et al. [G08] refer to *"the economics of cataloging web resources "* being different to that of cataloging books emphasis being on the interoperability of a MARC record across libraries as opposed to *"dynamic and more transient"* web resources where document access can be more direct through the use of coding [S00]. As a result metadata for web resources have been within the framework of coding schemes like the *Guidelines for Electronic Text Encoding and Interchange* and the *Standard Generalized Markup Language (SGML) "providing for the identification of document attributes when and as they occur in the machine readable text"* [S00]. The extensible markup language XML and the semantic

web with its resource description languages and vocabularies classically META Tags, Dublin Core, RDF, OAI and in principle aids for resource discovery on the internet represent the state of the art web resource markup and description . A selection of relevant interoperable vocabularies will be elaborated in section 2.3.

However, according to Gilliland et al. [G08] for digital library and archive resources, human-created metadata have retained the legacy as cataloging extending their controlled vocabularies to facilitate *"intra-community knowledge sharing"* [G08] The identification and cataloging of digital objects serves the time dependent management and archiving of the digital objects with structured metadata supporting organization and access. [GH00]. "All archives use some form of metadata for description, reuse, administration, and preservation of the archived object" [GH00] accompanied by resultant challenges related to the metadata creation, standards and structural rules in addition to the level at which the metadata are applied and the storage. The result is a variety of metadata formats depending on project type, data type, discipline and archiving rules and recommendations. This heterogeneous scenario and the need for intra-discipline knowledge sharing is well illustrated by the differing metadata descriptions classified according to their role as summarized by Gill et al. [G08]: descriptive data structure standards for a selection of resource descriptions e.g. MARC, Dublin Core, MODS, EAD

- markup languages and schemas for encoding metadata in machine-readable syntaxes e.g. XML and RDF
- Ontologies for semantic mediation between data standards e.g. CIDOC CRM
- Protocols for distributed search and metadata harvesting e.g. Z39.50, SOAP and OAI-PMH fig. 2.2 below illustrates an OAI Harvesting Model

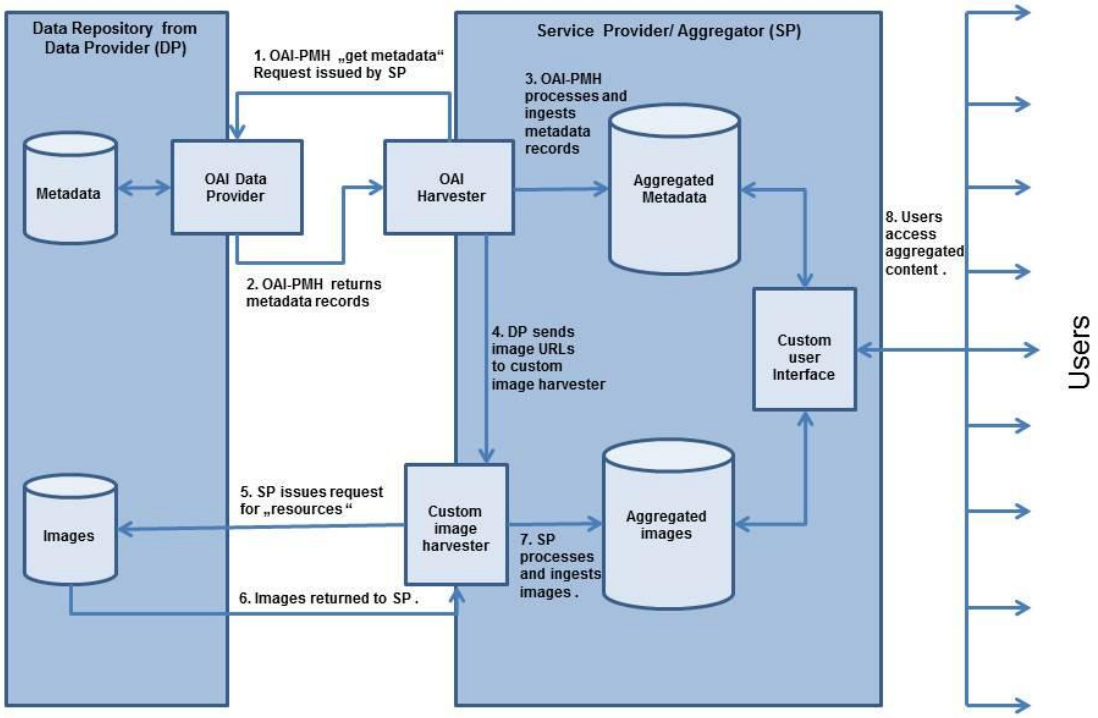


Fig. 2.2 Metadata Harvesting Model [QA10]

Description Principles

Chapter 2 has until here illustrated the state of the art on the basis of user expectations and bibliographic aspects as the basis for any metadata creation framework. The illustration has however also highlighted the need to specify the framework based on the bibliographic system and its contents i.e. the data elements and the bibliographic language. This section now looks at the guidelines for designing or in this case controlling the language in line with the basic idea behind the dissertation i.e. encapsulating encoding and data structuring aspects into a graphical user interface. These guidelines are specific to customized metadata vocabulary control and implement the description of archive content. This description effectively constitutes the structural framework in accordance with the bibliographic objectives upon which the graphical user interface will act. Generally, the specific description principles complement the design principles of sufficient reason and parsimony whereby the latter achieves algorithmic preference and are in effect summarized as follows [S00]:

- **user convenience**
  Focus should be on user oriented descriptions
  - ***common usage***
    Vocabulary should be normalized according to popularity
- **representation**
  "a description should be based on the way an information entity describes itself"
  - ***accuracy***
    Descriptions should aim towards a perfect depiction of the described entity
- **sufficiency and necessity**
  The goal is to achieve stated objectives and exclude data elements which are not required
  - ***significance***
    only metadata of bibliographic significance are to be included
- **standardization**
  Bring conformity at all possible levels and to all possible extents

- **integration**
  Description should be based on a common set of rules

These principles also form the basis for a summative evaluation of the bibliographic system and are together with a series of interface variables elaborated on, in chapter 6.

Summary

Chapter 2.1.2 summarizes the basic characteristics of the state of the art in bibliographic metadata creation frameworks. The fundamental message is that the foundations of sound digital archive frameworks lie in the more bibliographic description and vocabulary as opposed to the meta-language for tagging data elements. In other words, the metadata framework of any bibliographic entity can only be achieved by implementing a controlled vocabulary upon which a meta description language can be implemented and in so doing outlining the borders of the framework in question. With such a base in place one, one or more metadata tagging schemes can then be selected for structuring the data elements in question and realizing the “virtual online archive”. Now the goal of the dissertation is an abstraction of the tagging process and tagging language and therefore the next stage would be to implement the framework as a conceptual graphical user interface. The next section therefore looks at state of the art structural markup upon which the resulting bibliographic metadata framework can be implemented.

2.1.3 Metadata Encoding and Processing

In the field of humanities in which archiving is implemented as described in this dissertation, the concept of data encoding has always played an important role. The encoding tools have always been aligned to technological advances as a result of the proliferation of web technology these concepts have had to be integrated in new technological forms and older data sets transformed to comply with the new technological environment. This rapid technological development is among others the main motivation behind developing an abstract implementation framework for archivists given that their main area of interest is the collecting and archiving of documents and artifacts of interest as opposed to exploring optimal methods for data processing. This section looks at the state of the art of data encoding in general and the weighted metadata relevance in the

process, serving as the basis for structuring archive data and providing for the possibility of automated data processing and exchange.

Metadata in Digital Archives

Now having described metadata as structural and bibliographic markup, we now proceed to access metadata relevant to digital archives in general and digital manuscript archives such as the Jonas Cohn Archive in particular. The general idea is to single out aspects of interest which can be implemented within the metadata creation framework. In order to do so we need to look at the reasons behind the need to create the metadata and how these reasons weigh in, in comparison to each other. These “reasons” can be described in accordance with the tasks the metadata carry out and are summarized below as follows:

- Archive Description

  By looking at the title of the dissertation it is obvious that not all archives are the same and that not archive management tasks involve managing the same thing. In other words archive management tools should distinguish between physical archives and digital archives, artifact archives and text archives, printed text archives and archives containing scripts.

- Machine Readable Records

  The common term for retrievable information stored as records on audiovisual and computer media where machine based equipment is necessary for reading the information held by the records. In this dissertation archive records refer to descriptive information or facts relating to objects physical or otherwise representing the archive material.

- Encoding Digital Objects

  Compound digital resources can be referred to as digital objects characterized by their fundamental description elements i.e. bibliographic data, file and object type. The objects can be any integrated compound digital resource e.g. article, photo, record, journal.

- Encoding Text

The standardized structured representation of texts and their description in digital form enabling the machine-readability of the texts concerned. Text encoding is central to the dissertation for texts that cannot be read by character recognition hence needing both transcription and object records with concurrent object and text encoding. Resulting object heterogeneity and duplication of tasks is to be resolved by the framework and its mapping facility.

- Encoding Finding Aids
  Traditionally archive and library records serve as finding aids for locating physical objects contained in the archive/library. In the digital case metadata assume the same role highlighting and describing archive collections. Finding aids and description activities are summarized under the umbrella term and standard Encoded Archival Description EAD.

- Context Description
  Where interoperability and collaborative work is concerned aspects of the context around the archive encoding activities are not to be missed. These aspects including primarily user roles, data capturing and ownership issues and are incorporated in Archival Context Description.

Integrated Digital Objects

Digitization scholars and archivists render physical archives in their digital context mapping predefined structures of existing conceptual organizations in a World Wide Web domain. As a result metadata and their encoding serve to store, process and present archive contents in this domain. The role and objectives of and behind metadata and their encoding lie within the annotation and markup of integrated digital objects classified as:

- *content* - information contained by document
- *structure* – content location arrangement
- *presentation* – rendering content and structural information

Marked-up components of the three basic classes mentioned above constitute the backbone of an SGML/XML based digital document referred to in literature as the three

layer distinction. This three layer distinction aims to separate the user interface from archive data [WP99] or content structure from format whilst describing structure and semantics. Hillesund et al. [TH02] do not totally agree with this concept and subscribe instead to the idea of *integrated XML markup or encoding*. Hereby, semantic markup or content-oriented markup is bundled up with the respective descriptive elements under the common term presentation markup whilst structural markup defines the notions of logical and hierarchical structuring [TH02]. The general understanding is that of a heterogeneous conglomerate of marked up XML data serving information delivery upon processing.

Heterogeneity

The concept of heterogeneity in general and heterogeneous metadata for digital archives in particular introduced above can be defined within the framework of digital object structure and presentation introduced in the preceding subsection. Di Lorio [AD07] defines document level heterogeneity on a document segmentation basis i.e. document content and structure are the same and can be considered segments despite different presentation medium. On the other hand Hillesund et al. [TH02] describes heterogeneity from a publisher's view with single input to distributed publications as illustrated in fig. 2.2 emphasizing the integrated and *“interwoven”* [TH02] nature of encoded documents. Integrated digital archives reflect both views encompassing encoded content intermixed with graphic images of the content with the presentation as the *“most evident part”* of the digitized archive. In other words heterogeneous metadata encoding reflects upon single input – multiple output encoding activities involving content-oriented and structural markup elements presented across a single web-based medium. Subdividing the encoding process results in the following categories:

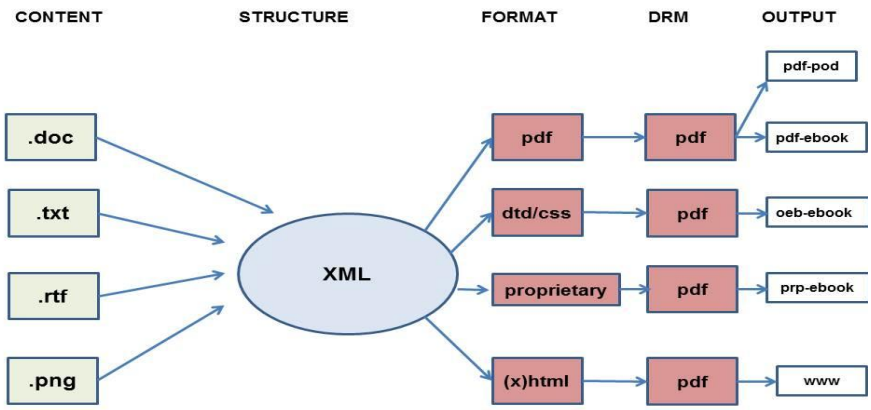


Fig. 2.3: Hillesund's heterogeneity [TH02]

- structural markup
- semantic markup
- presentational markup

Semantic Heterogeneity

Marked-up online presentations of digital archives are generally part and parcel of the semantic web community. Subsequently, interoperability, multiple documents and objects i.e. heterogeneous content reflect on a common domain however on the basis of different data schema and resemble the phenomenon of semantic heterogeneity, a phenomenon common whenever there is more than one data structuring standard. The reconciliation of the semantic structured or semi-structured data builds the backbone for data exchange and interoperability and is hence of importance to any data creation framework. Common to literature on semantic heterogeneity is the notion of *"semantic mappings"* or *"mediation"* with reference to transformation expressions specifying data *crosswalks.* Novel to the concept of semantic heterogeneity in this dissertation is more the notion data creation as opposed to retrieval and hence more towards the idea of a single source multiple schema frameworks. However, the need for heterogeneous semantic metadata reconciliation remains decisive for interoperability activities despite the single source multiple schema architecture. On the other hand the notion of *data heterogeneity* is a phenomenon associated with description tags, in this case differently named tags refer to the same data

elements. In the case of digital archiving this phenomena is dealt with on the basis of standardized references *"Verweisungsform"* and is therefore not subject to assessment.

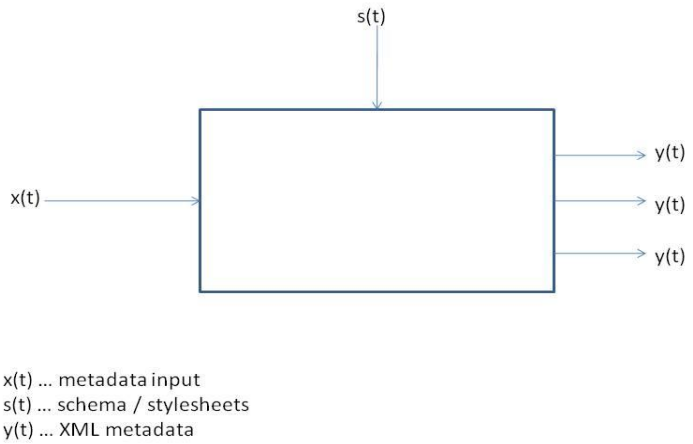


Fig.2.4: Overlapping Metadata

The motivational scholarly archive to this dissertation, the Jonas Cohn Archive, consists mainly of handwritten manuscripts written in the standard handwriting of the author's time. The integrated digital objects of the archive are hence heterogeneous with photographic images of letters and diaries and their bibliographic summaries delivering the same information content via different media in the same domain. Furthermore, text and catalogue resource description metadata describe the content and semantic structure of digitized objects whilst metadata object description and transmission encoding describe the format and presentation of the same information on a digital object level. In other words graphic elements intermixed with content and presentation and containing the same structures and information as XML fragment describing the graphic contents i.e. heterogeneity. Fig. 2.4 above summarizes the abstract characteristics of metadata overlap across multiple XML vocabularies and the notion of avoiding multiple data capturing of elements containing the same content. This resembles the core feature of my proposed framework outline in terms of single input multiple output XML description as illustrated in Fig. 2.5 below

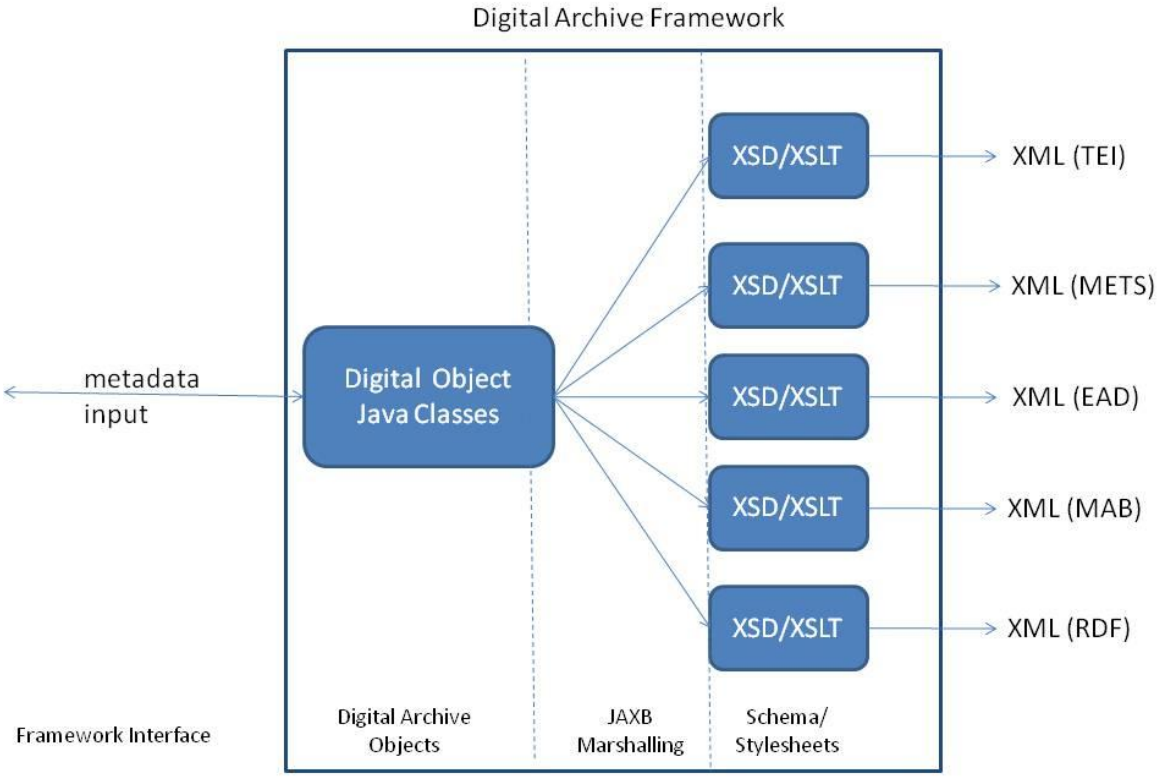


Fig.2.5 Heterogeneous Metadata Creation Framework System – digital objects java classes elaborated in chapter 4 XML Binding

Ontology Mediation

Formally speaking the notion of ontology is aligned to knowledge representation defining associated concepts and their relationships. However, this representation is in effect a specification of structural frameworks for information organization as the basis for data integration, information finding as well as data interchange purposes. As such *ontology mediation* is concerned with ontology reconciliation for design reuse and heterogeneous scenarios, Bruijn et al. [BE06] subcategorize it to include *ontology mapping*, *ontology alignment* and *ontology merging*. A formal description of the definition given by Bruijn et al. [BE06] summarizes ontology mediation as *"the reconciliation of differences between ontologies in order to enable interoperability"*. Mediating overlapping metadata as shown in Fig. 2.5 above resembles *ontology alignment* and enables interoperability.

Metadata Relevance

Looking at the encoding process illustrated above reveals the diversity and extensibility of XML metadata in digital archives hence the need for metadata specification in relation to their relevance. The classification of digital archive objects mentioned in the preceding subsection indirectly defines the metadata relevant to a digital archive within the framework of structural and textual and bibliographic markup languages as described in section 2.1.2. In our case these metadata are heterogeneous in nature as they describe text and image objects delivering the same content. Scholarly archives such as the Jonas Cohn Archive contain handwritten material, bibliographic finding aids for navigation and identification as well as text summaries reviewing archive content.

Digital archiving renders digital object classes to text and images describing and illustrating the content, hence relevant metadata include:

- Bibliographic metadata
  Structuring and presenting archive records and literature
- Object collection metadata
  Describing and structuring ordered digitized pages
- Encoded text metadata
  Structuring text reviews and correspondence

The relevant metadata serve as the basis for interoperability in addition to their structuring of non-optical character recognition readable texts in image form consolidating the integrated nature of the compound digital class objects. Original pages can then be presented in book or lecture form and navigated as real objects. The relevance of selected metadata will be weighed in and assessed according to the following application scenarios:

- Collection Catalogues
  Archivists and librarians are primarily concerned with collecting and cataloguing appropriate resources and maintaining normalized collection records. Respectively, metadata focus is on digital record keeping and resembles classical record cards and supporting centralized record collection. In other words, the XML encoding and description of finding and structuring aids.

- Digitized Objects

  Handwritten and other texts not transcribed remain text material for the user and must be presented as such i.e. as *"an ordered hierarchy of content objects"* [AD07]. Hence metadata describing text structure and order applied to digital images of text in the presentation context e.g. METS/MODS. Furthermore, presentation plays an important role regarding the usability and acceptance of digital archive hence Di Lorio's description of presentation as *"effects the reader's recognition of content objectives… derived from medium used to access content"* [AD07].

- Automated Data Storage

  Network activities and data interchange amongst archiving and encoding institutions via grid storage systems and repositories e.g. Kalliope, Textgrid or DARIAH. Whilst the infrastructure for multiple storage and centralized searching is now in place thanks to the aforementioned initiatives, the framework of this dissertation focuses on the actual creation of the metadata serving as input for them. In other words, the framework covers the existing gap between infrastructure and digital archives assisting archivists by the automated generation of structured documents in formats compatible with the preservation storages and centralized search infrastructure. Research therefore focuses on integrating appropriate compatible collection metadata in the compound digital manuscript collection archive and interoperability with other networks.

2.1.4 Metadata Abstraction and Patterns

Abstraction

The principles of abstraction and patterns are common from object oriented software development aiming to accomplish an *"unambiguous description of a data structure"* [AG05] by specifying the *"services"* of the data structure. Both principles render reusability with the former facilitating data structure reuse for all cases requiring the same set of *"services"*. Abstraction refers therefore to the modeling of sets of similar objects with their associated set of common services.

On the other hand according to relevant literature among others Pardi et al. [WP99] the principle characteristic of XML is data description and structuring on the backbone of the capability to separate user interface and data i.e. a form of abstraction. Applying both principles to the concept of a single input multiple output metadata framework discussed in section 2.1.3 and fig. 2.6 & fig.2.7 the separation of the user interface and XML data as well as common metadata and the respective XML schema resembles an abstraction of the metadata to be collected using this framework. By modeling the framework as such, the archivists do away with duplicated XML schema based metadata encoding and the metadata to be collected can be modeled as object classes. In other words, metadata are *encapsulated* within the archive framework with schema based outputs as the processed outputs presented to the importing archive or collection. In so doing the principles of separation of content, format and presentation can be achieved whilst utilizing concepts common to object oriented software development.

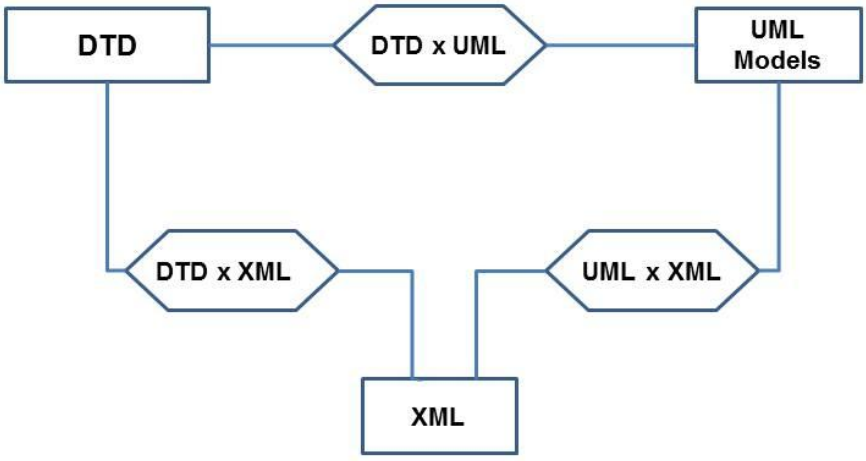


Fig 2.6 UML - XML Transformation Framework [AV02]

Patterns

The notion of identifying cooperating classes incorporating an extensible framework relies on the distinction between application domains. This distinction dictates the common metadata relevant to the application domain in question. Heterogeneity as specified in this work reflects upon common metadata descriptions of heterogeneous digital objects across heterogeneous XML schema. From an abstraction point of view, these metadata resemble a recurring design and can be identified and specified by software design patterns which identify pertinent reusable objects specified as classes [EG95]. Such patterns ease *"reuse of architectural knowledge and artifacts"* providing *"a common vocabulary and shared understanding for design concepts"* [BS07].

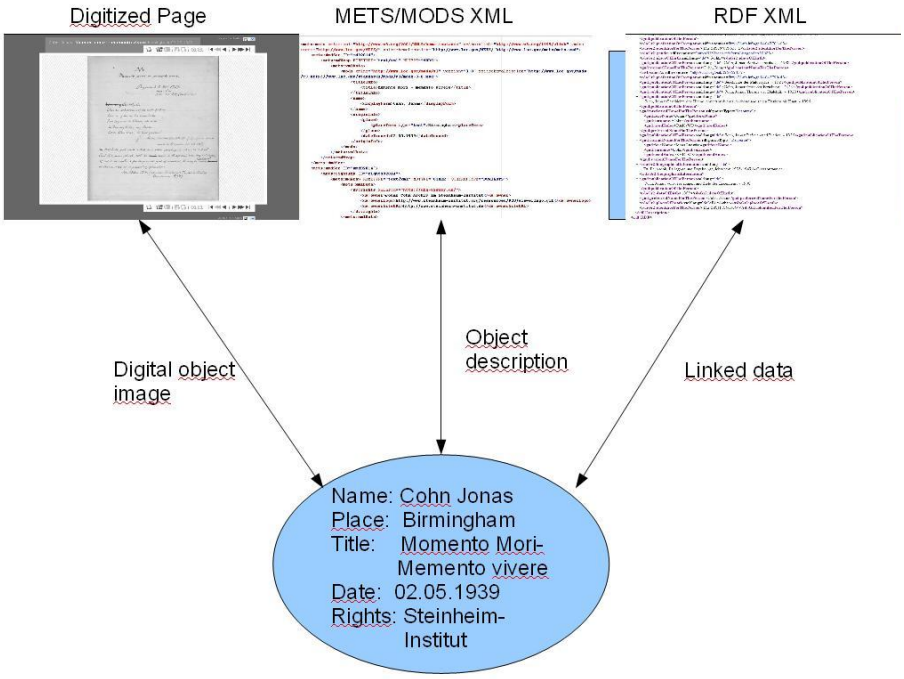


Fig. 2.7*:* Data overlap in multiple object semantic heterogeneous metadata scenario

In the diagram above the common metadata *Name*; *Place*; *Date*; *Rights* and *Title* are presented via three different media i.e.

- Digitized image
- XML METS/MODS fragment of digitized image
- XML RDF linked data library record

The diagram resembles a heterogeneous scenario with recurring design aspects and reusable objects such as author's name, place and rights. Modeling metadata elements as reusable objects enables the identification of design patterns and normalizes the data capturing activities within the encoding process. These objects can be designed applying object oriented principles, implemented as classes with respective attributes and relationships which in turn can be marshaled to appropriate XML formats hence simplifying the metadata creation process. This transformation is illustrated by the pattern in Fig. 2.8. Gamma et al. specify objects in general as consisting of data and procedures i.e. methods and operations stimulated by an interaction, in this case a client request [EG95]. Relevant pattern types and their description are summarized below:

Pattern Classification

Generally patterns are classically characterized as creational, structural or behavioral according to purpose. Creational patterns concern object creation; examples of structural ones are listed below.

Structural Patterns

- Bridge
  decouples an abstraction from its implementation
- Façade
  caters for a higher level interface accessing and simplifying subsystem use
- Composite
  allows uniform treatment of objects and their compositions by clients with the objects composed into tree structures to represent hierarchies
- Flyweight
  efficient fine-grained object support through sharing
- Decorator
  provides dynamic extended functionality alternative to subclassing
- Adapter
  resolves interface incompatibility aspects for interacting classes [EG95]

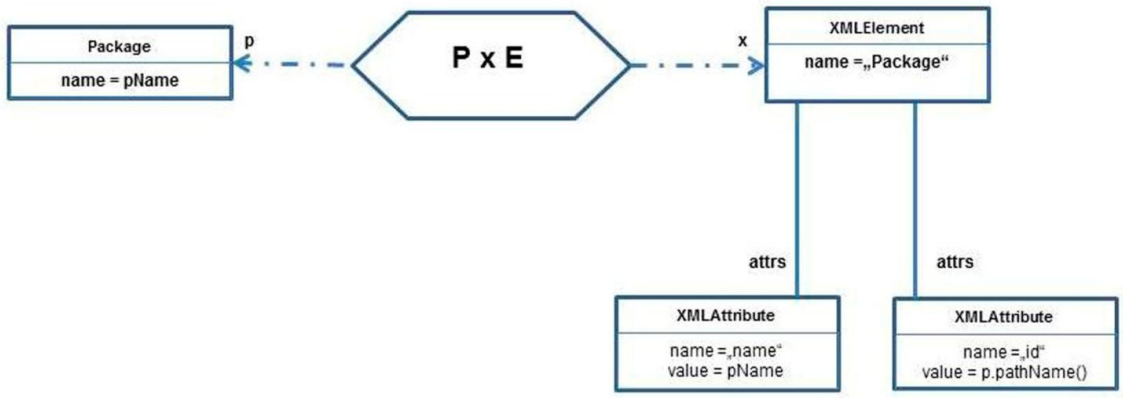


Fig. 2.8 UML to XML transformation using patterns [AV02]

Semantic Web Design Patterns

- Users Add Value
  architecture allows implicit and explicit user data entry
- Network Effects By Default
  “sets inclusive defaults for aggregating user data as a side-effect of their use of the application” [BS07]
- The Long Tail
  “leverage customer-self service and algorithmic data management to reach out to the entire web” [BS07
- Cooperate
  pattern architectures offering a *“network of cooperating data services”* [BS07]:
    - Web services
    - Interfaces
    - Content syndication
    - Data reuse Services
- Perpetual Beta
  design considers applications *“ongoing services”* as opposed to software artifacts and hence engages users as *“real-time testers”* with feedback flowing into new feature design [BS07]

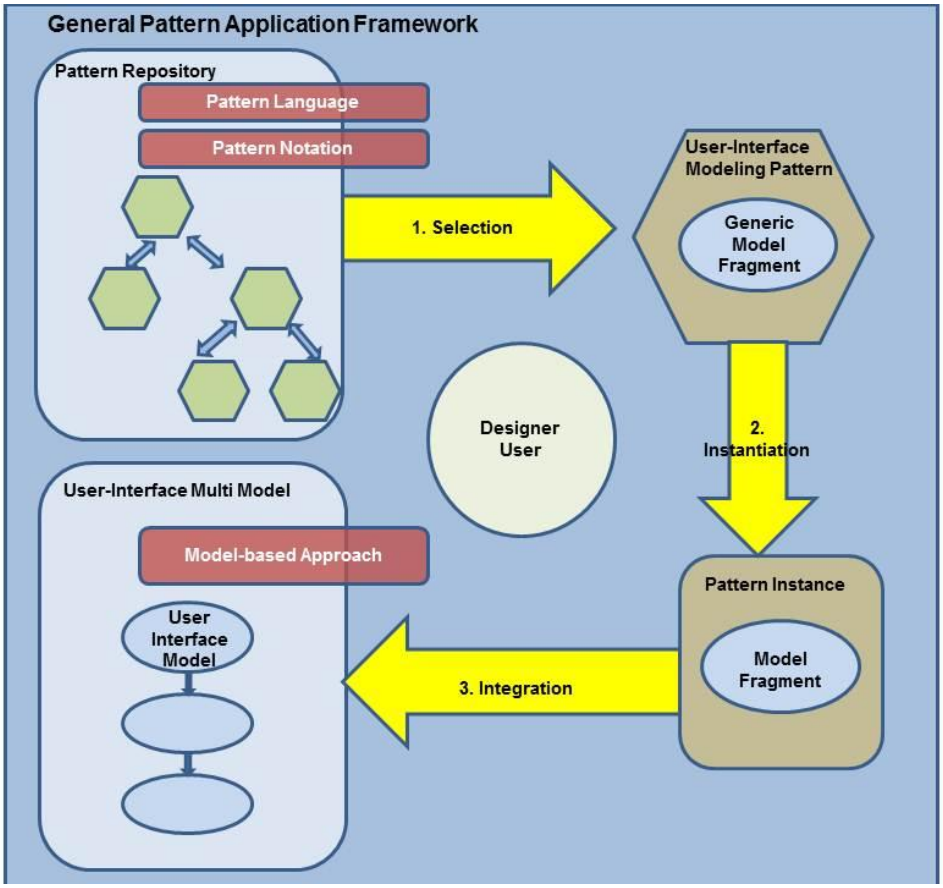


Fig 2.9 Pattern Application Framework [RF07]

Interface Design Patterns

- Polyvalent-Program Pattern
  characterizes applications with an architecture allowing polyvalent i.e. multiple interfaces
- Filter Pattern
  specific to execution of a non-interactive program which processes standard input and produces a standard output
- Cantrip Pattern
  specifies a non-interactive status generator with no input and no output
- Source Pattern
  describes data emitting non-interactive applications with no input
- Sink Pattern
  describes non-interactive applications that only accept input data without an output

- Compiler Pattern
  characterizes non-interactive file transformation programs
- Separated Engine and Interface Pattern
  specifies an architecture separating the application core logic from the presentation i.e. interface and user interaction. *"The engine and interface roles are normally realized as separate processes"* [BS07]

Task Patterns

Some overall concept of patterns in software development in general have been summarized above, however limitations to internal system design similar to those for object oriented design are evident. Paterno [P99] proposes the notion of *Task Patterns* as design support for activity oriented user interface design characterized by the following specifications:

- **P1** Pattern name
- **P2** Problem addressed by pattern
- **P3** Task relationship specification
- **P4** Specification of objects manipulated by task
- **P5** Scenario of use
- **P6** Possible sub-patterns
- **P7** Aspects that can be modified in an instance
- **P8** Applications where it is likely to be used

Figure 2.10 illustrates a Task Pattern modeled along the Paterno's Concur task tree notation and specifying a search task. In addition to addressing the search task problem, the pattern further specifies tasks relationships, the objects manipulated and related scenarios as characterized by the notion of Task Patterns summarized above.

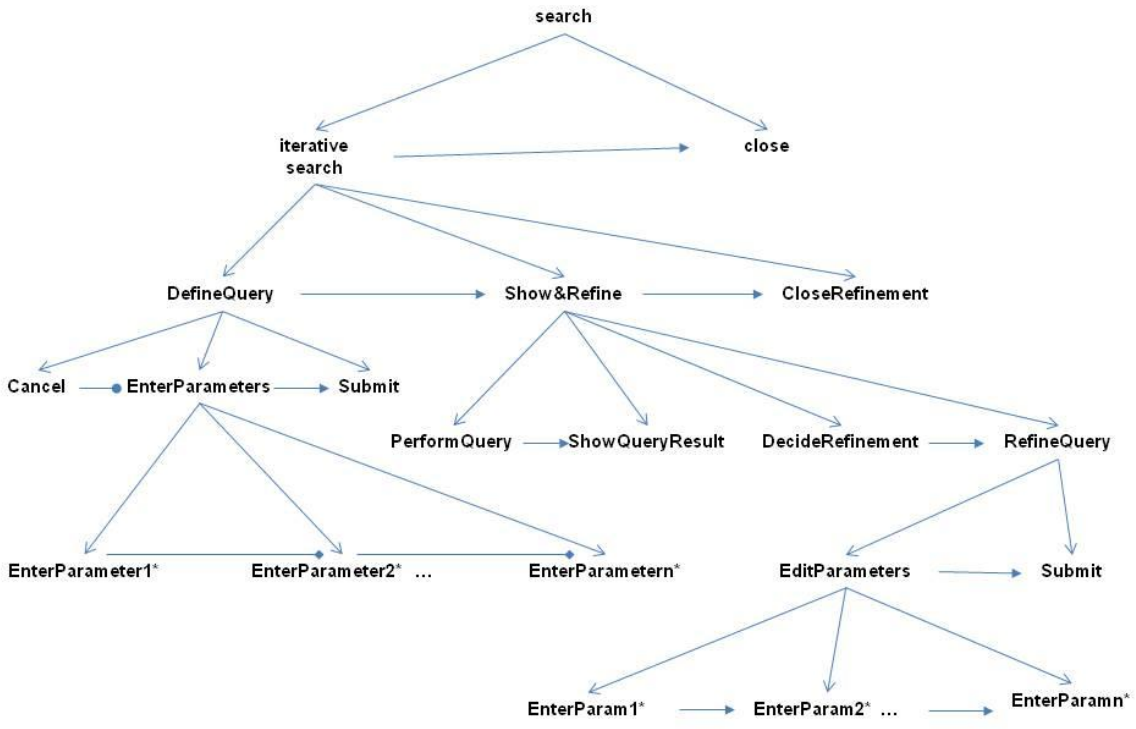


Fig. 2.10 Search Pattern Specification [P99]

Patterns in the Metadata Creation Framework

The notion of patterns introduced above plays an intermediary role in the design and implementation of the framework and its graphical user elaborated in the succeeding as well as in the analysis and design of the user tasks in the metadata creation process. Whereas the latter dealt with in chapter 3 influences the user tasks and the usability of the entire framework, the former help model the architectural and system design aspects of the framework whilst considering software reuse and maintenance. Further specific architectural patterns to be introduced in chapter 4.2 influence the presentation and control of the metadata record collection process. This process includes the modeling of the collection interface together with associated abstractions, the presentation as either audio and visual components or display and format view components and lastly the functional implementation as control components. In other words patterns influence the framework at system level enabling a comprehensive and manageable framework system and architecture.

At the content level patterns in general and semantic patterns in particular help single out vocabulary and metadata overlaps across the semantic heterogeneity illustrated by the XML schema standards for digital archives. Here patterns bring in semantic independence with metadata and elements being abstractly described at object and hence entity level. This abstract description encompasses the core hypothesis of the dissertation harboring metadata content as abstract objects accessed via an interface with automated generation of the structured XML document. In other words patterns serve the abstraction of the archive metadata as a milestone towards schema independent metadata creation and the associated encoding as sought by the framework's problem description.

Summary

In summary, the metadata creation framework introduced in this dissertation applies techniques common to computing sciences and software engineering to the field of metadata encoding and in so doing supporting the idea of *"text & presentation as integral parts of document identity"* [AD07].

Due to the heterogeneous nature of the required metadata, encoding is both document-centric with objects as *"a rendition of format and presentation"* [AD07] as well as data-centric serving structural and automated processing requirements. Furthermore, interoperability aspects imply heterogeneous interoperability serving subsections of the integrated domain community. Encapsulation and abstraction are to assist the object oriented normalization process whilst patterns ensures schema neutrality and design reuse on the data capturing level i.e. the data level as well as interaction level represented by the user interface.

As such hiding encoding implementation and representation in an object *encapsulation* and customizing the arrangement of the objects and classes to be implemented *pattern* within a set of cooperating classes *framework* accessed via a set of requests to which the digital objects can respond *interface* [EG95] constitute the abstract encoding framework.

Subsequently, in the succeeding subsection we will summarize state-of-the-art metadata frameworks in both the digital library science and digital archiving disciplines looking in particular, at their capability to cater for integrated metadata, heterogeneity and interoperability. In addition to that, the subsequent subsections of chapter 2 are dedicated to XML encoding of the metadata itself commencing with XML as a meta-language, metadata types, structured data schemes and the tags describing the metadata. An excursion to digital editions and cross media publishing completes this section on the state-of-the-art in metadata encoding.

2.1.5 Digital Metadata Frameworks

The notion of digital metadata framework draws upon the concept of digitization under the auspices of cultural and digital preservation. The umbrella term digitization encompasses a multitude of activities associated with a set of archival artifacts which in most of the cases are in text form. These activities may include online publication commonly referred to as “digital editions”, preprocessing textual data for automated further processing a refinement of which cross publishing may be the result. Cross publishing refers to the single input based processing of textual artifacts for publication as both online and hence digital editions as well as traditional text editions in book form. The processing is considered heterogeneous [TH02] is described in the preceding sections of chapter 2.1.3 together with the notion of heterogeneity in metadata and their vocabularies. State of the art digitization subscribes indeed to the common notions mentioned above however focus being on standardization and interoperability in the spirit of the open archives initiative referred to in chapter 2.1.2 and the idea of common access to information. Standardization follows interoperability as it is the prerequisite for the basis of a common understanding i.e. interoperability between digital archive require a common language.

Furthermore, with digital archive aiming to utilize the worldwide web as the infrastructure for proliferation and communication of archival contents; the dictation of a common gateway and language becomes imperative.

To this effect the eXtensible markup language has crystallized itself as the common gateway language and hence resembles the foundation upon which standardization and consequently interoperability are derived. Now because digital archives are not confined to a particular subject or subject areas, standardization on the basis of an extensible markup language results in a multitude of standards all being applied to the same corpora, a phenomena referred to as semantic heterogeneity and also dealt with in the preceding chapter 2.1.3. This phenomena and the encoding of XML metadata in such a scenario resembles the problem analyzed in this dissertation. Whereas current solutions focus on the tedious concept of crosswalking using extensible stylesheet language transformation (*XSLT*) scripts the hypothesis of this dissertation and hence the alternative sees a structural abstraction of the metadata encoding and subsequent mappings of the metadata with a selected standard culminating in a standardized XML document.

The hypothesis and the surrounding heterogeneity concepts having being introduced as part of the preceding sections of chapter 2.1.3. As such digital metadata frameworks can be seen as mediating tools accommodating the automation of the XML metadata creation process as part of the standardized structured digital archiving process common to cultural heritage preservation and proliferation activities. The archives dealt with in this dissertation are mostly based in the humanities and as in the case of the Jonas Cohn Archive containing handwritten or non-machine readable contents whose digitization involves the use of text images as conservatory measures. An example of such a text image is illustrated in the figure below.

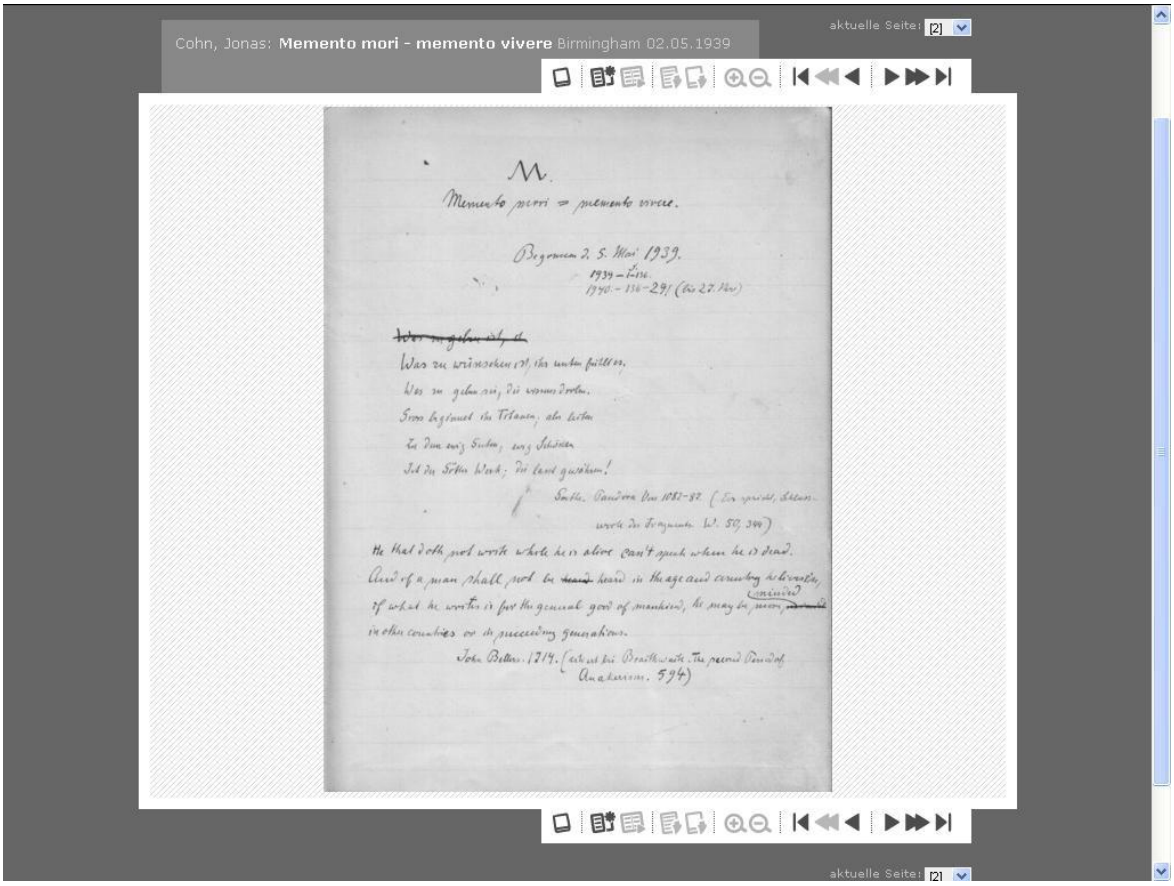


Fig. 2.11 Digitized METS/MODS encoded text image of Jonas Cohn's Memento Mori handwritten manuscript

An Introduction to the XML Meta-Language

The extensible markup language XML is defined by its developers the W3C as "a simple, very flexible text format" [W3C] originally designed for large scale electronic publishing but now enjoying importance in data exchange on the internet. The introduction on structured mark-up in section 1.2 expands upon this role, exposing XML as the *state-of-the-art* meta-language for encoding electronic documents facilitating data interchange and automated processing whilst inheriting the role of preserving cultural heritage. Furthermore, with structured data interchange XML facilitates interoperability between standardized systems whose semiotics provide a platform for comparison via the definition of the XML tag sets. As an extensible generalized markup language [WP99] XML characterizes a meta-language whose tag sets describe either itself or any other document structure specified via the Document Type Definition DTD. Pardi et al. [WP99] attribute XML vocabularies to this meta-language characteristic which further distinguishes XML as a

"profile" or *"subset"* of the pioneering Standard Generalized Markup Language *(SGML)* as opposed to being an application of SGML as is the case with HTML. In other words, the property of being a meta-language enables the development of further XML based description vocabularies and languages for marking up data and thereby opening additional application areas for XML. The result is specialized language specifications for hyperlinking schemes (*XLL*), for stylesheet languages (*XSL*), transformations (*XSLT*) and schema definitions (*XSD*). These language variations enable XML use for a further series of purposes to include:

- XML as a data interchange format
- XML for web data
- XML for creating common data stores

XML Characteristics

An insight into the basics of XML as a meta-language has been introduced in the previous subsection, with the implication that structured XML is defined as being concerned with *"describing information"* [WP99] and that XML documents containing such descriptions can then be used for data storage, interoperability and web data purposes. As a result XML resembles SGML in the sense that applications of XML in general and HTML in particular compliment this meta-language to display the information or data it contains. Pardi et al. [WP99] speak of XML's structuring and describing of data which is in turn formatted and displayed by HTML on the internet. It is therefore logical that we now go on to look at the characteristics of such an XML document and the relating processing in light of the complementary relationship with HTML. The XML as such represents the product of the metadata creation framework with its contents describing the content, context and presentation of heterogeneous archival artifacts in heterogeneous semantic formats. Generally an XML document is associated with the following characteristics:

- **Declaration**

  In order to be identified as an XML document, each document commences with a declarative tag known as the *"processing instructions"* or "XML declaration". This tag serves to deliver information on the version, encoding and structural status i.e.

definition of constraints

<mets:metsxmlns:xsi=http://www.w3.org/2001/XMLShema-instance xmlns:xlink="http://www.w3.org/199
:mets=http://www.loc.gov/METS/ xsi:schemaLocation=http://www.loc.gov/METS/ http:www.loc.gov/mets… ">

Fig. 2.12:XML-Declaration & Processing Instructions for METS

- **Elements and Attributes**

  In general, XML is hierarchically structured in line with its compatibility with the Document Object Model (*DOM*) as a mark-up language for web data and hence illustrates a tree structure. An XML element is then a tag set consisting of an opening and closing tag culminating with the document tag which acts as the root of the document of concern. Further elements of the XML are then nested in the document element setting up a *"parent/child relationship"* with the root i.e. document element. This notion of nesting also common to all hierarchical structures is defined as *"the process of embedding one object or construct within another"* [WP99]. Nesting in XML documents can be further extended to include embedding other XML documents in an XML document. However the general case of nesting XML elements confines the definition to include all child elements with the exception of the document element as being entirely resident within the document element. Elements may contain additional informational information outside the data subset in the form of attributes. These attributes are composed of quoted attribute names and values as part of an opening or empty tag element. The attributes themselves are as in any other programming languages not extensible as they resemble "data about data" and hence are descriptive in nature resulting in their incapability to contain multiple values and tree structures. An example of the attributes of an element is illustrated below.

- **Entities**

  In the subsection above the characteristics of XML as meta-language describing and structuring content are described. However, a further associated attribute of XML illustrated by the structuring process is the data storage and transfer characteristic. As such the W3C definition of an XML document refers to *"one or more storage units"* [W3C] in which the content is stored, and these units

represent XML entities with the *"document entity"* being the exception of not containing content. Entities may be classified as being either general or parameter entities depending on their use with the latter being parsed and intended for use in Document Type Definitions (DTD) and the former exclusively within the document content.

The preceding processing aspect leads us to further categorization of XML entities as parsed or unparsed respectively referring to their content as either *"replacement text invoked by entity references"* or any resource other than XML in text form or otherwise *"invoked by name using entity references"* [W3C]. The notion of namespaces also mentioned above is elaborated in the subsequent subsection on namespaces.

- **Document Type Definition DTD**

  XML documents are further characterized by their encoding with respect to their having a correct syntax and where applicable the validity of the code against a defined constraint. In the former case the documents are characterized as being well formed whereas in the latter case characterization refers to validity. In this scenario a document type definition DTD dictates the structure of a *"well formed"* XML document setting out a set of elements on the basis of which the document is declared "valid" upon successfully fulfilling the criteria laid out by this element set. The DTD is contained in an external file and referenced using the document type tag <! DOCTYPE>

- **XML Schema**

  A further XML structural constraint consisting of an ensemble of tags describing the structure of an XML document is the XML Schema Document *(XSD)* commonly referred to as simply an XML schema. This structural constraint in the XSD file has proven popular as an infrastructure for defining and validating the correctness of content created in a digital archiving domain. Further advantages include automated data manipulation and processing in addition to easier definition and identification of data patterns within XML. Together with the definition of data facets these characteristics of (standardized) XML schema and their XSD files

enables the integration of databases as persistency in addition to the easier transformation between different schema and standards i.e. cross-walking, concepts of which form the basis of the framework developed for this dissertation as metadata creation novelty.

- **XML Namespaces**

  As an extensible mark-up language, element names in XML are open to definition resulting in the need for a *"name conflict"* [W3C] mechanism of which XML Namespaces resemble a simple method for doing so. In this mechanism XML names are classified as being either prefixed or unprefixed names and defined within a namespace as qualified names which can be correctly interpreted once the namespace is referenced in an XML document. The XML Namespace is then identified by a uniform resource identifier and declared as namespace binders by a family of XML attributes commencing with either the element **xmlns** or **xmlns:** directly or by default. The Metadata Encoding and Transmission Schema declaration (METS) illustrated in the figure 7 above had its namespace declared as a schema instance together with the linking namespace for XLink. The scope of influence of the declared namespace stretches between the opening and the corresponding closing tags applying to all elements and entities encompassed within this scope.

Processing XML

With XML having been specified as an extensible structural format delivered as a mark-up language with the intention of enabling information interchange and data structuring, the specification of how the structural format is used to operate on and produce XML documents is left to the other XML specification constraints such as XSLT, XML Schema and XQuery. It is these XML based constraint languages which *"describe and specify the processing relationships between XML resources"* [W3C] within the framework of the XML Processing Model. The processing model describes an interoperable method for illustrating the sequence in which processes should be applied to XML documents. The interoperable processing constraints applicable to digital archives are listed below and give insight to the processes necessary to create XML documents and hence providing for their accommodation within the metadata creation framework.

Extensible Stylesheet Language (XSL)

The principle notion which makes XML very attractive for data structuring and interoperability is the separation of the data from the presentation elaborated on in the succeeding subsection. To this notion belongs to the concept of processing data with XML and the presentation, processing and associated stylesheets with XSL and XSL transformations. Hence XSL as a set of recommendations is composed of the constraint languages, XSL Transformations (XSLT), XML Path language XPath and a "*vocabulary for specifying formatting semantics*" XSL-FO [W3C]. The former play a significant role in creating XML metadata for digital archives and is therefore summarized below:

XSLT

The XSL Transformation language XSLT is a semantic set of elements which uses the stylesheet language XSL to describe the transformation of XML documents into other XML documents and (X)HTML and hence serving the presentation and display of XML in a web-browser. In the digital archiving domain XSLT has played a major role when cross walking metadata records across the semantic heterogeneity of the description languages standards represented by XML Schema and further meta-language oriented structural constraints. The transformation itself is expressed as a valid, well formed XML document with the validity of names declared in the XSLT Namespace referenced using the following URI:

```
http://www.w3.org/1999/XSL/Transform
```

Fig. 2.13 XML-Declaration & Transformation Instruction

An XSLT transformation document is considered a stylesheet set upon a template or pattern containing rules which serve as a basis for matching with the hierarchical elements of the XML document. The template rule can be implemented as a pattern in the source tree or an instantiated template as part of the result tree. The processing of XSLT elements is forward compatible allowing the import and combination of one or more stylesheets in addition to their embedding as either an XML or non-XML resource. In

general XSLT is implemented on source, result and stylesheet XML documents whilst processing is in within the framework of the XML Processing Model.

The applicable data model is also shared with XPath and implemented on the XML documents as hierarchical trees encompassing a set of seven node categories summarized as follows [W3C]:

- Root nodes
- Element nodes
- Text nodes
- Attribute nodes
- Namespace nodes
- Processing instruction nodes
- Comment nodes

XPath

XPath is a finding aid recommendation which addresses navigation in XML documents and defines parts of the XML document using XPath expressions and an ensemble of built-in functions. XPath uses non-XML syntax to support the manipulation of strings, numbers and Boolean operating on the abstract logical structure of the hierarchical XML document. With XPath expressions as the primary syntactical construction XPath can match XML nodes to selected patterns therefore shared functionality with XSLT and XPointer, the former of which has been introduced in the preceding subsection.

Separating User Interface and Data

As a meta-language XML is concerned with making statements about data and hence all about *"describing information"* [WP99]. The describing text is characterized as ***Markup*** and *"all that is not Markup constitutes the **character data**"* [W3C] resulting in the separation of the user interface from the data [WP99]. The self-description aspect emanates from the necessity of XML documents to conform to a set of rules. Parda et al. interpret this self-description as the ability to provide metadata *"so that the data in the*

documents can stand apart from the formatting that describes how the document is displayed" [WP99].

This notion is criticized and at the same time utilized by Hillesund [TH02] as the basis for the doctrine of *"one input –many outputs"* from which the single input encoding concept described for the dissertation question emanates. The basic idea behind this *"doctrine"* is that of the separation of content and structure in XML. This separation is the foundation of the notions of separating XML data and the interface as well the document structure and meaning as propagated by Pardi [WP99] as mentioned above and in Hillesund's criticism towards Harold et al.'s claim [TH02]. The general conclusion to the debate is that XML describes *"structure and semantics but not formatting"* [TH02] where the formatting is in reference to the document's appearance. In the digital archiving field the appearance of the XML structured artefacts is only of concern to either the interface or the hypertext presentation and hence beyond the scope of the archival structuring process.

As such a plausible approach to abstract multiple structure metadata encoding is one where a centralized structured XML encoding is hosted in a framework the access to which and the presentation of the contents is left to a separate user interface also hosted by the framework. This resembles the proposed solution to the dissertation question described in detail in the succeeding chapters. The structured encoding is concerned with the semantics and standards to the digital archive is subjected to whilst the graphical user interface provides for interaction with user both for the standardized structuring and the presentation of the archive contents.

2.2 Metadata in Digital Archives

In digital archiving the concept of metadata revolves around interdisciplinary activities in computing and the humanities focusing on hypertext web technology and classical library and information sciences to disseminate and or provide access to stored archival material. In other words, the notion of metadata in this context serves the purposes of informing whoever it may concern of the contents of any described archive digital or otherwise with the assistance of descriptive information which we then call metadata. The information contained within the metadata can be faster accessed when these are organised in categories making it easier for target users to focus only on those categories of interest

and hence reducing complexity and effort. To this effect, modern computing provides a marvellous and widely accessible infrastructure via the internet in addition to tried and proven algorithms accelerating such searching activities. At the same time data structuring concepts and algorithms available via modern computing further facilitate preservation activities supporting network oriented grid multiple storage which provide backup and information security in the case of system errors or any catastrophe potentially resulting in the loss of data in general. In summary, the notion of metadata refers to categories of descriptive and structural information providing insight into the contents, context and composition of the entities they describe whilst preserving access and assisting with the management of any collection of these entities. In the case of state-of-the-art digital archives, these categories and the information they contain are represented by specified XML tags understood by a community of users to convey a particular structures whilst reflecting the content, context and composition of the archives. Although this definition refers to real objects and documents, it is not confined to these and extends to include web documents meant for publication on the internet. With the internet being a worldwide web of documents accessible via a common infrastructure, this infrastructure dictates a document description understood by all users.

This document description resembles metadata within the web domain describing not only the structure of a web document but also the content, context and composition enabling easier and accelerated access to the document via any access point in the infrastructure. To elaborate on this issue we shall now look at metadata based on their purpose and the limits of this purpose within a digital archiving environment. To this end we will describe metadata as elements of library information sciences, digital editions of topics of research humanities and as descriptive elements of the presentation media represented by the internet and the worldwide web.

2.2.1 Metadata Types

In general, metadata types and standards for digital archives are developed and defined by national libraries and archives and foremost by the library of congress in the United States and *“Deutsche National Bibliothek”* [DNB] in Germany the former specifying descriptions for local digital archives. In the same manner description entities and hence metadata for worldwide-web documents are defined and developed within the framework

of the worldwide web consortium W3C and its respective specifications and recommendations [W3C]. As a result metadata types are respectively classified according to their function in the bibliographic and online presentation domains. This aspect resembles a refinement of the structural requirements in digital archiving now specifying the structures concerned and their relation to the content and presentational needs of the archives. The implementation of these structural requirements then takes place at the schema level where the individual functions are translated into metadata vocabulary and hence description tag elements. These vocabularies and their associated tag libraries then enable machine readability and automated processing and are describing in further detail in the succeeding subsections of this chapter.

In addition to the classification of the metadata type categories we go on to look at the individual functions from which the element vocabularies are derived.

Bibliographic Metadata Types

In the case of bibliographic metadata, the functionality classes of the metadata types resemble the management tasks involved when running a digital archive. In addition to this, these categories also take into consideration the overall aims of digital archiving namely digital preservation and automated processing resulting in the following subcategories of metadata types:

- **Administrative Metadata**

  The process of managing digital archives and hence creating metadata is a collaborative activity by nature as it involves multiple users cooperating to create a product entity; only in this case the time frame involved in this creation process is continuous. As such part of the archive management tasks and the resulting metadata for their description is concerned with the creation and ownership of the digitized artefacts and is referred to as the administrative metadata. Administrative metadata are applicable both as bibliographic metadata as well as semantic resource descriptions either as administrative or creator elements. In addition to information on the creation and intellectual rights, elements of administrative metadata further describe provenience and storage of the digital contents of the

archive in addition to themselves being either embedded within the descriptive element documents of the objects or externally serving as reference metadata.

- **Descriptive Metadata**
  This type of metadata serves to describe an archival object either on the basis of archival or bibliographic description categories set by standardization body or defined by the digital archiving task requirements. The descriptions can be defined as modules of the object classes they describe. Classical descriptive metadata include XML fragments in Text Encoding Initiative TEI, Encoded Archival Description and Metadata Object Description.

  - TEI Manuscript Description
    TEI provides for the encoding infrastructure assisting description activities of the digital versions of archived manuscripts.

    The digital version may either be a transcription of the original manuscript or the ensemble of digital images all nested within the manuscript description element <*msDesc*> [TEI09]. The *msDesc* element hosts the description of a *"single identifiable manuscript or other text bearing object"* [TEI09] complemented by the following set of components:

    - Manuscript identifier <*msIdentifier*> containing indexing information
    - Heading <*head*> of any type e.g. title, glossary etc.
    - Manuscript contents <*msContents*> hosts the *"intellectual content"* [TEI09] in a paragraph series form or as *"a series of structured items"* [TEI09]
    - Physical description <*physDesc*> part or full description optionally further complemented by descriptions of the model.physDescPart class.
    - <*history*> *"group elements describing the full history of a manuscript or manuscript part"* [TEI09]
    - <*additional*> assembles additional information on the manuscript and its surrogate copies integrating bibliographic information with

the curatorial, provenance or administrative information.

- Manuscript part <*msPart*> relating information on a previously distinct manuscript now resembling a part of a composite manuscript. [TEI09]

- MODS Digital Document Description

  As opposed to defining description elements for each possible archival artefact and its respective digital object, MODS possess a set of Top Level elements and attributes used throughout the schema. The differentiation of object types and classification categories is handled by the <*typeOfResource*> MODS element. Although the Top-Level elements span across the entire MODS Schema, a selection of these elements overlap with description elements from other standards. These elements include:

  - Identifier <*identifier*>
  - Title Info <*titleInfo*>
  - Identifier <*identifier*>
  - Physical description <*physicalDescription*>
  - Extension <*extension*>
  - Part <*part*>

  The remaining fourteen elements are dealt with in the succeeding subsection on the Metadata Object Description Schema in their elaborated form.

- Encoded Archival Descriptions

  The descriptive aspects of metadata encoded in EAD are concerned with description aids aimed at simplifying search activities hence the name finding aid encoding. The target documents are therefore solely for this

purpose and hence the encoding infrastructure is classified in general, linking and tabular display attributes

- **Structural Metadata**

  The notion of structural metadata describes the structural aspects from the presentation and navigation point of view. In other words the reference to structure points towards the hierarchical architecture which dictates the flow of navigation and the arrangement of the digital objects in the presentation. Whilst on one hand the navigational aspects relate to the nested sub-elements as in any hierarchical architecture. On the other hand the presentational structure may require the implementation of pointers and structural links as in the case for structural metadata in METS. Here the hierarchical structure of the encoded object is described using the structural map element <*structMap*> which in turn contains the nested attributes multiple METS pointer <*mptr*> and file pointers <*fptr*>for identifying related content.

2.2.2 Semantics and Resource Description Metadata

Andleigh et al.'s [AG05] definition of a semantic network focuses on the *"graphical representation of a relationship between two objects"* as a directed graph with four primary nodes i.e. *concepts*, *events, characteristics* and *value*. The relational nature justifies the concept of semantic networks for relational data sets or databases as well as in class based structural relationships. On the digital archiving side Hillesund et al. [TH02] refer to "*semantic markup or content-oriented markup"* as the scenario where meta-language elements such as XML elements describe the content of the elements *"often using humanly understandable element type names"* [TH02] In light of this one acknowledges the aspect of semantic heterogeneity dealt with in chapter 2.1.3 and the associated reference to semantics in relation to the structuring and descriptive XML Schema based description vocabularies.

An analysis of the definitions of semantics introduced above reminds us of the interdisciplinary nature of digital archiving in both the structural preservation and the structural presentation sense. Whereas Andleigh et al. [AG05] define semantics from a computing and semantic web point of view, Hillesund's [TH02] view of semantics is more computer philological. With the presentation and the preservation success of the concept of digital archives being attributed to the internet and associate computing technologies these view require unification for an optimal use and propagation of the archive contents.

To this extent the notion of representing relationships between archival artefacts is gaining popularity in digital humanities and as a result semantic metadata encoding in the web computing sense is on the increase. It is therefore no surprise that a number of national archives including the German National Library have begun to offer their MARC records in the semantic web format Resource Description Framework RDF. To illustrate this phenomenon I have chosen a handwritten manuscript page in Fig.2.14 digitized for the Jonas Cohn Archive. The digital version of this document is encoded in METS/MODS as part of a digital manuscript of the original diary hosted by the archive. The author Jonas Cohn and his respective works are also encoded in MARC in the central national archives for preservation and presented as MARC 21 metadata accessible as MARCXML online. The digital text image of a manuscript page in the digital version of the Jonas Cohn Archive is illustrated in the figure below:

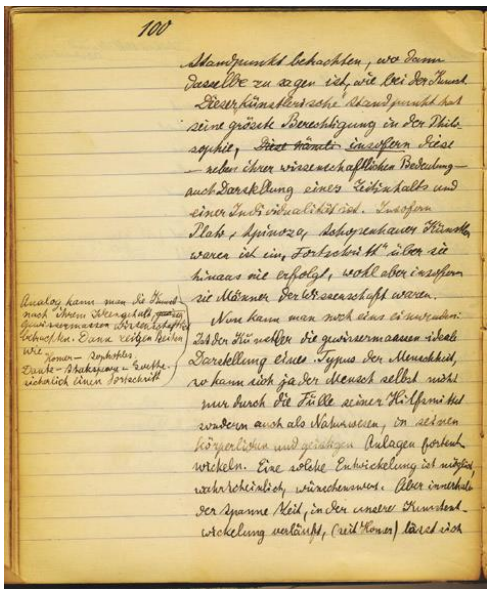


Fig. 2.14 Digitized Text Image from the Jonas Cohn Archive

This digital object of the image in Fig.2.14 above is encoded in METS/MODS as part of Jonas Cohn's academic diaries and may be linked to his national archive entry now available in RDF. The respective author entry as MARC21 bibliographic entry is shown below in addition to the complimentary XML fragment for semantic and resource description purposes linking the author of the manuscript to his registered intellectual products in digital form. It is also worth noticing that this *Person* record is closely associated with the indexing Personen Normen Datei number, PND element common only to German records and often subject of international element vocabulary discussions. The PND and its GND variation for institutions indexes owners and hosts for intellectual artefacts providing for name variations and aggregations.

Link zu diesem Datensatz: http://d-nb.info/gnd/118669664

Person: Cohn, Jonas (männlich)

Andere Namen: Cohn, Jonas; Cohn, Jonas Ludwig (vollständigere Namensform)

Quelle: M; Biogr. H Emigr.

Lebensdaten: 1869-1947

Beruf (e): Psychologe; Philosoph

Land: Deutschland (XA-DE); Großbritannien (XA-GB)

Weitere Angaben: Dt. Philosoph, Pädagoge und Psychologe; lebte von 1933-1947 in Großbritannien

Sachgebiet(e): 4.7p Personen zur Philosophie; 5.5p Personen zur Psychologie; 6.4p Personen zum Bildungswesen

Fig. 2.15 Jonas Cohn's DNB PND entry

The description fragment illustrated in the figure below shows the XML Encoded resource descriptions of the PND entry shown in the figure above. The fragment is encoded in Resource Description Framework format for the semantic web with bibliographic elements describing and identifying GND and hence MARC21 defined bibliographic content.

<rdf:RDF>

<rdf:Description rdf:about=http://d-nb.info/gnd/118669664>

<rdaGr2:identifierForThePerson> (DLC)n 83063515</ rdaGr2:identifierForThePerson>

<gnd:countryCodeForThePerson>XA-DE</ gnd:countryCodeForThePerson>

< gnd:variantNameForThePerson>Cohn, Jonas Ludwig</ gnd:variantNameForThePerson>

<gnd:publicationOfThePerson xml:lang="de">Cohn, Jonas: Religion und Kulturwerte – 1914

</ gnd:publicationOfThePerson>

<foaf:page rdf:resource=http://de.wikipedia.org/wiki/Jonas_Cohn/>

< gnd:preferredNameForThePerson rdf:parseType="Resource">

< gnd:foreName>Jonas</ gnd:foreName>

< gnd:usedRules>RAK-WB</ gnd:usedRules>

< gnd:foreName>Jonas</ gnd:foreName>

< /gnd:preferredNameForThePerson>

<rdaGr2:dateOfDeath> 1947</ rdaGr2:dateOfDeath>

<rdaGr2:professionOrOccupation rdf:resource=http://d-nb.info/gnd/4047701-0/>

<rdaGr2:identifierForThePerson> (DE-588c)4089531-2</ rdaGr2:identifierForThePerson>

<rdaGr2:gender rdf:resource=http:/RDVocab.info/termList/gender/1002/>

Fig 2.16. PND entry in XML RDF format

Resource Description Framework Metadata

A general summary of metadata and more specifically metadata in digital archives would make reference to the set of descriptive vocabularies illustrated by markup elements to give meaning to the content of the body being described. This summary of metadata leaves us with the following classifications from a metadata point of view:

- *Extracted metadata*: obtained automatically from document contents.
- *Explicit metadata*: determined by humans upon analysis of the documents in question
- Metadata for resource discovery
- Metadata for rights management and access control
- Metadata for administration and preservation.

Having now acknowledged the need for more than a bibliographic and archival description vocabulary, we will now look at semantic web description vocabularies adding on to the above mentioned metadata sets to include semantic web metadata. These semantic web metadata have their own XML vocabularies and structuring along the recommendations of the Resource Description Framework [W3C].

Introduced in chapter 1.2 on structured encoding and structured markup, the Resource Description Framework is finding application as part of the semantic web description of bibliographic and archival metadata as illustrated above. This description language follows the axiom of *"metadata is data"* [W3C] assuming the role of a *"first class object"* for the purposes of describing, accompanying or embedding within another document. This axiom is complemented to *"metadata can describe metadata"* to result in the reference to *"machine understandable information about web resources or other things"* [W3C].

As this dissertation is mainly concerned with the encoding of explicit metadata for digital archiving and preservation, the RDF Schema is presented only as summary of the element classes as illustrated in the figure below.

Class name	comment
rdfs:Resource	The class resource, everything.
rdfs:Literal	The class of literal values, e.g. textual strings and integers.
rdf:XMLLiteral	The class of XML literals values.
rdfs:Class	The class of classes.
rdf:Property	The class of RDF properties.
rdfs:Datatype	The class of RDF datatypes.
rdf:Statement	The class of RDF statements.
rdf:Bag	The class of unordered containers.
rdf:Seq	The class of ordered containers.
rdf:Alt	The class of containers of alternatives.
rdfs:Container	The class of RDF containers.
rdfs:ContainerMembershipProperty	The class of container membership properties, rdf:_1, rdf:_2, ..., all of which are sub-properties of 'member'.
rdf:List	The class of RDF Lists.

Table 2.2 RDF class names

Summary

The subsection above shows the associated concepts behind the semantics term in the digital archiving sphere. Whereas semantics in the computer linguistic sense encompasses all content descriptions and humanly understandable description elements, the semantic web approach has the W3C's machine understandable web resource information. The goals of digital archiving introduced in chapter 1, however combines the two notions to include the machine readability and the *"classical"* structured encoding giving meaning to structured content.

Abstracted Data

The principle behind metadata draws upon the notion of information encoding, systematically representing information as messages with sets of descriptive words, letters, or signs generally accepted and understood by a particular group of people. In the context of digital archiving, the term metadata generally refers to structural and descriptive data for managing and presenting content—hence, the complementary reference to "data about content", which alludes to the "data about data" reference. This data about content can be structural elements (defining a document's structure) or descriptive elements (application data illustrating the data content). As such, the first step towards standardized schema-independent abstract encoding reflects the semiology of the descriptive metadata elements represented by the elementary bibliographic elements and their attributes, which constitute the structure and description of a digital archive. Therefore, key aspects

include the identification and modularization of abstract standardized elements to contain the metadata records.

The identification process may be guided by the digital archiving recommendations and reference models of national or regional archiving bodies, such as the Regeln für Nachlässe und Archive (RNA) [WK10] or the Recordkeeping Metadata Schema (RKMS). Such recommendations and reference models are usually compatible with or reflect international metadata standards, such as the Functional Requirements for Bibliographic Records or Dublin Core, as is the case with RKMS. Abstraction and schema independence illustrate the modularization of metadata elements culminating in a standardization of the metadata and their descriptions. The representation of this standardization in reference models such as the Open Archives Initiative System (OAIS) and the SPIRT Recordkeeping Metadata Schema provides a platform for optimal crosswalking and metadata interoperability. The resulting module element descriptions resemble an algebraic illustration of the intersection of relevant heterogeneous schema library elements.

Metadata abstraction does not override the structure, data model or functionality of the descriptive elements meant to contain the archival records; instead, it makes them schema and syntax neutral, thus avoiding several problems associated with hard coding, including obsolescence. The difference to other object-relationship approaches, such as that of the RKMS, lies in the fact that we do not envisage a further XML metadata schema as a framework for mapping the metadata, opting instead to utilize existing schemata. One example of such a schema is Encoded Archival Description (EAD). EAD is a data-structuring standard which contributes to the modularization of shared data types by correlating descriptive areas with content standards. However, EAD focuses on publicly accessible data structures (as opposed to collection management) whilst, at the same time, neglecting presentation content.

In the mediated heterogeneous metadata encoding process presented in this article, data abstraction is part of the broader concept, whereby archival descriptions are created outside the XML spectrum using other technologies—in this case, object orientation, but also relational databases. Furthermore, data abstraction is encompassed in the data model, shunning the definition of further XML description tag libraries as modules in

favour of classical language and syntax-independent, integrated, model-based abstractions. The actual metadata encoding in XML takes place within the mediating unit examined in Section 2.3 as the population of structured container elements. The resultant metadata model specified by the abstraction is therefore characterized by the following:

- Standardized element library
- Object-oriented database
- Schema independence

2.2.3 Metadata Creation Tools

AllegroHans:

The metadata and resource description objectives outlined in the preceding subsections are characterized by the respective metadata creation activities. These activities be they cataloguing, text encoding or simple tagging are in most cases supported by and reliant upon software tools as the backbone for the metadata creation framework. In the case of the Jonas Cohn Digitization Project prototype for this dissertation work, the recommended tool in the context of the DFG Archiving Recommendations [DF10] is AllegroHans [AH97]. The origins of the allegroHans tool lie in the cataloguing world of mainly libraries in German speaking countries. AllegroHans resembles a derivate in the common version Allegro-C developed University of Braunschweig Library [AC80]. AllegroHans is an open source product however platform specific with enterprise support provided for members of the user consortium. Implementation of the software outside the context of the consortium required bilateral consultations with the developers. Advantages included the RNA customized encoding structure along recognized digitization and encoding recommendations [WK10] as well as the functional requirements for bibliographic records [FR08]. Metadata interchange and registration with central archiving authorities is on a bilateral basis without automated data interchange. On the other hand data collection and interchange structures in AllegroHans and allegro-C marked-up in Unicode [AH97] with unstructured export to XML reflected an obstacle to state-of-the-art preservation, obsolescence and presentation. The developers offered, in the context of the Jonas Cohn project custom enterprise solutions on a bilateral basis to overcome the challenges faced.

In addition to the aforementioned structural challenges the allegroHans tool's focus on bibliographic catalogues and records resulted in meagre provisions for heterogeneous metadata creation involving images and multimedia in general and object relational associations in particular. Due to the Unicode structuring and missing structured XML capabilities, state-of-the-art structural vocabularies could not be accommodated. Furthermore, due to missing obsolescent structural vocabularies accommodating all digital objects beyond text associated to an artifact, state-of-the-art metadata interchange and propagation media could not be utilized.

Greenstone:

A further tool for building and for the propagation of digital library collections is the Greenstone suite. The focus of the Greenstone suite lies in the building of digital libraries and the suite enjoys prominent propagation and exposure as it is a development of the UNESCO and the Human Info NGO [GL00]. In addition to being open source and having a global distribution, the suite is platform independent and runs on most conventional operating systems. Interoperability is on the basis of the Open Archives Protocol for Metadata Harvesting (OAI-PMH) [QA10] and data formats predefined. Predefined format is the Dublin Core [DC09] in addition to further formats with origins in the Asian-Pacific area. Interoperability in general and data export and ingestion in particular exists with respect to the Metadata Encoding and Transmission Standard METS [ME10]. Professional customizing of the internal metadata formats can carried out using Greenstone's Metadata Set Editor and data ingestion with assistance of plug-ins is available among others for MARC, OAI, METS. In the case of digital archives, the size of the Jonas Cohn Archive, the Greenstone tool resembles a very complex and specialized tool whose usability requires a dedicated encoding archivist. The ingestion and export concept deals primarily with structured heterogeneous metadata and is therefore unsuitable for single input heterogeneous metadata creation. The prerequisite for ingestion and export remains structured metadata, the missing point in projects in which structured metadata are to be created. As a complex tool with dedicated interfaces for user and librarian streamlined across languages, formats and purpose, Greenstones reflects usability problems for archivists in the academic sphere as their primary concern lies in the archive content and not the dissemination. As such this tool proved too

complicated and difficult to handle for the target users to deal with, within the context of this dissertation.

Archivists Toolkit:

Described as an open source archival data management system, this tool was *"intended for a wide range of archival repositories"* [AT06]. The focus lies in the archival processing, the production of access instruments in addition to the promotion of standardized metadata structures aligned with efficiency and low cost training in the perceived archiving process. The system is capable of describing archiving material in the Encoded Archival Description (EAD) [EAD09] format and the export features including MARCXML, METS and Dublin Core exports. However, the focus on US standards resulted in questionable compatibility with local archiving rules and recommendations. As such metadata creation infrastructures resembled interoperable standards; however structural vocabularies and associated interfaces in the context of the RNA [WK10] could not be confirmed. In addition to this, language barriers especially on the interfaces and the encoding of semantic notation common in German speaking areas saw a trend towards a negative trend of acceptance, naturally in the context of the Jonas Cohn digitization project.

Summary

Looking at the selection of tools as illustrated above revealed the need for a home grown customized system suited to the needs of the target community. Whereas the local frameworks deal adequately with the categorization structure and semantic requirements of archives in general, they lacked sufficient integration of XML technology and the object relational aspect associating digital objects with their descriptions and resources across the spectrum of digital media. On the other hand the set of frameworks supporting standardized international formats and requirements declined to serve local recommendations and interoperability issues. As such illustrating the necessity of a metadata creation framework addressing object oriented digital archiving issues whilst respecting interoperability and archiving recommendations, structural and preservation standards within the context of scholarly academic archives.

2.3 Structured Data Schemes

2.3.1 Dublin Core

The term Dublin Core Metadata refers to interoperable metadata standards developed by the Dublin Core Initiative which describe metadata and defines them as „*data about data*“ articulating „*a context for objects of interest*“ [DC09]. The definition is extended further to “*a formal description of inner and outer characteristics of traditional and digital documents and objects that supports their availability to the public*” [KA03]. The standard thanks its popularity to idea of “*core metadata* for simple and generic resource description” which comprises of 15 elements providing interoperability within the framework of Open Archives Initiative Protocol for Metadata Harvesting.ISO Standard 15836:2009 [DC09].The context of use has focused on DC's role in a semantic worldwide web, enabling linked data movement together with the Resource Description Framework RDF, resulting in DC effectively being “a tool for cross-database searching” [KA03] within the context of semantics as illustrated in Fig. 2.17. The figure shows cross-database searching across the BIBSYS, Mavis and SEPIADES databases [KA03].

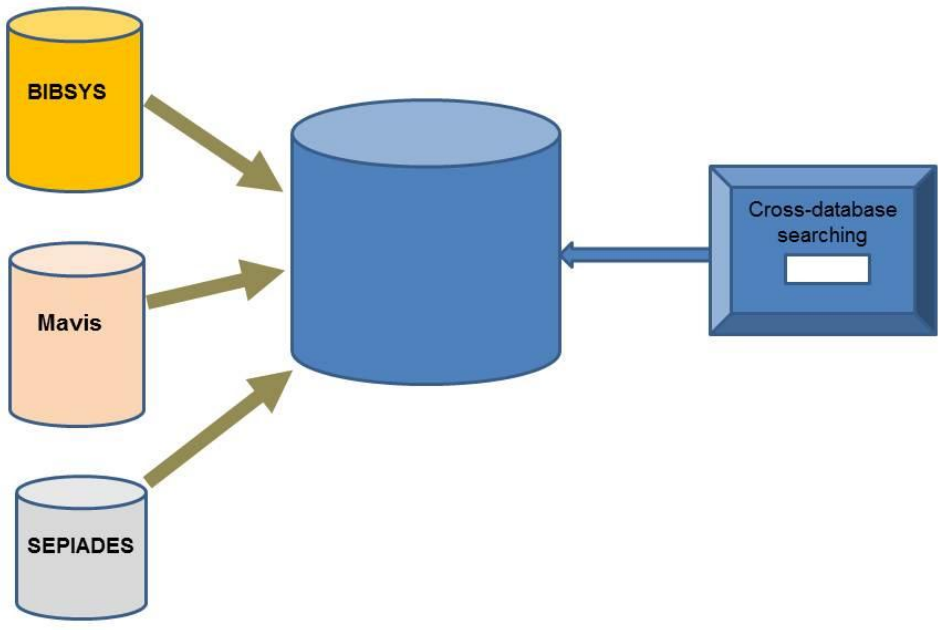


Fig. 2.17 DC a tool for cross-database searching [KA03]

Core Element Set

The Dublin Core Metadata Initiative entertains a wide range of resource description metadata vocabularies and technical specifications based on the following 15 core characteristics known as the “Dublin Core:”

- *title:* a name given to a resource
- *creator*: an entity primarily responsible for making the content of the resource
- *subject*: a topic of the content of the resource
- *description:* an account of the content of the resource
- *publisher:* an entity responsible for making the resource available
- *contributor:* an entity responsible for making contributions to the content of the resource
- *date:* a date of event in the lifecycle of the resource
- *type:* the nature or genre of the content of the resource
- *format:* the physical or digital manifestation of the resource
- *identifier:* an ambiguous reference to the resource within a given context
- *source:* a reference to a resource from which the present resource is derived
- *language:* a language of the intellectual content of the resource
- *relation:* a reference to a related resource
- *coverage:* the extent or scope of the content of the resource
- *rights:* information about rights held in and over the resource

Levels of Interoperability

Given the focus on web semantics and cross-database interaction, it is only natural that the core focuses on “application profiles” characterized by the following four levels of operability:

- *Level 1 Shared term definitions*

at this level metadata using applications interoperate on the basis of shared natural language definitions. The metadata terms are set by the respective application environment and “hard-wired” using individual implementation technologies. Interoperability outside the scope of the environment is not of priority.

- *Level 2 Formal Semantic Interoperability*
  RDF provides a shared formal model supporting Linked-Data promoting interoperability amongst the metadata using applications.

- *Level 3 Description Set syntactic interoperability*
  provides the platform for metadata validation and exchange based on the notion of Descriptions and Description sets as bounded entities with a specified identity comparable with a manageable record.

- *Level 4 Description Set profile interoperability*
  on the basis of a specification of formal constraints which provide an information model and XML expression of the constraints on a Description set leveled against the Description Set Profiles, DC-DSP.

  Figure 2.18 illustrates the abstract model of the Dublin Core Metadata Initiative for metadata interoperability as described in the preceding subsections. Notable is the entity relation of the description elements to the record object and the subsequent breakdown into the description entities, description, statement and value. The value string encompasses the syntax of the encoding scheme in relation to its vocabulary.

DCMI Abstract Model

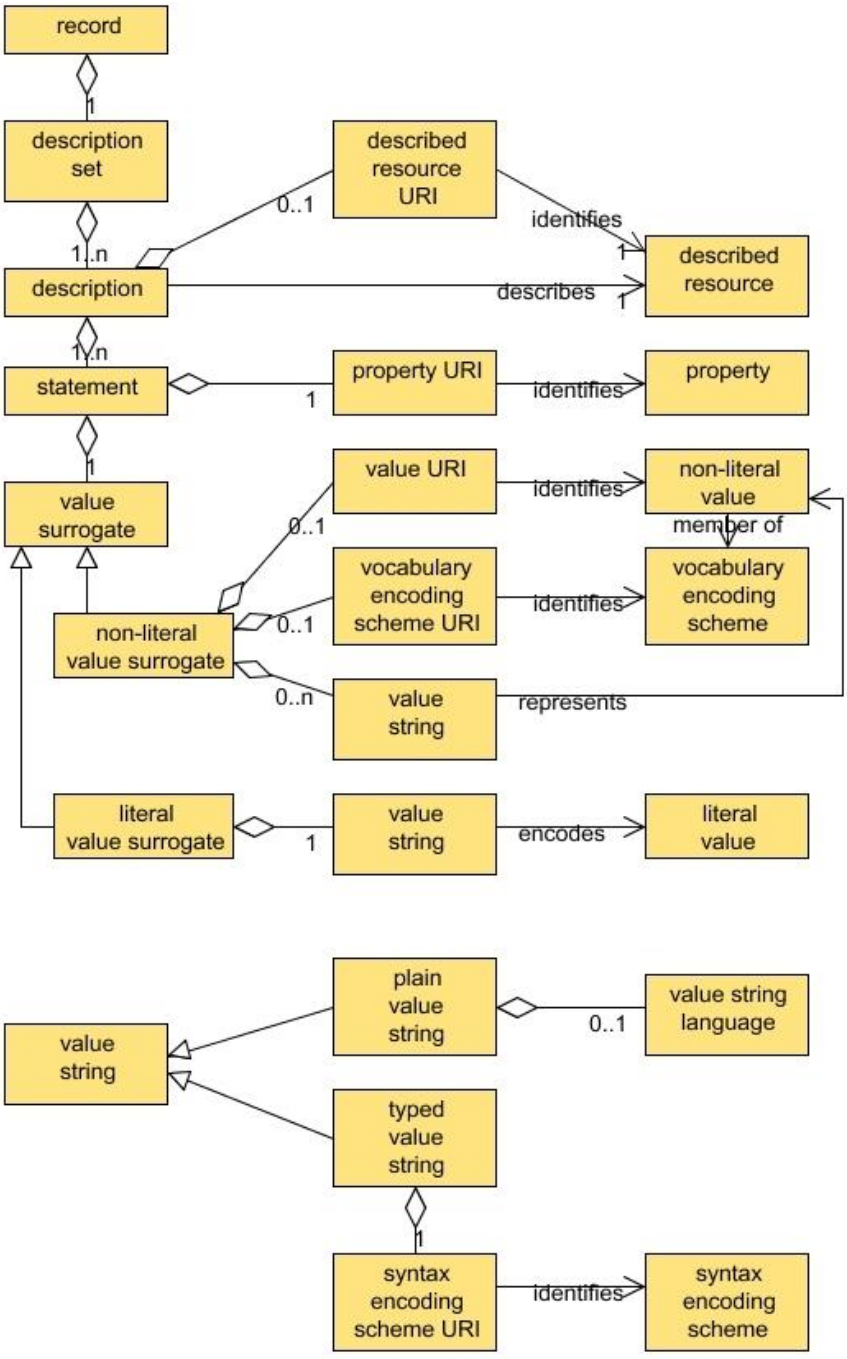


Fig. 2.18 DCMI Abstract Model [DC09]

2.3.2 Metadata Encoding and Transmission Standard

This metadata encoding standard belongs to the family of bibliographic library standards developed and maintained by the library of congress. Its main purpose is to serve the encoding of object metadata classified as descriptive, administrative or structural using XML syntax. Emphasis is on the necessity for successful management not only of physical material but also the digital objects and their use, hence relating the administrative to the structural metadata. Presentation aspects have the archive user in mind supporting user reflection of the technical metadata to the original document digitized for virtual usage. METS is often associated with the Metadata Object Description Standard also from the same family of encoding standards and together they form the basis of the DFG-Viewer [DV10] common for the presentation of digitized manuscripts. The standard resembles a compound description of structured digitized archive artefacts. METS documents assume are interoperable and assume an information submission (SIP), an archival information package (AIP) or a dissemination information package role in the OAIS Reference Model. METS documents are composed of seven main sections:

- ***METS Header <metsHdr>***

  metadata elements of this category describe content information about the objects within a document and the document itself. This information includes creation dates and editing agents including relevant attributes. Standard elements include:

  - <metsHdr>
    - CREATEDATE and RECORDSTATUS being the relevant attributes
  - <agent>
    - ROLE and TYPE attributes being relevant

- ***Descriptive Metadata <dmdSec>***

  The metadata elements in this section generally point to either internally or externally embedded metadata using the <mdWrap> and the <mdRef> metadata elements respectively. All <dmdSec> elements must possess an ID attribute for

unique indexing of each element providing for the linkage of sections of the metadata to selected parts of the digitized object:

- <mdWrap>
    - As long as metadata wrapped in this element are Base64 encoded and wrapped within <binData>, they can either be in XML syntax or be arbitrary binary or text.
- <mdRef> describes and contains a resource identifier (URI) pointer towards an external retrievable metadata source. The associated attributes include:
    - *LOCTYPE*: a specification of the locator type wrapped by the element

valid values
URL
URN
PURL
HANDLE
DOI
OTHER

Table 2.3 <dmdSec> Locator types

    - *MIMETYPE:*
      a specification element referencing external content-type descriptive metadata
    - *MDTYPE:*
      a description element illustrating the referenced metadata's form of which the following are among others valid

valid values
MARC
MODS
EAD
TEIHeader
DC
OTHER

Table 2.4 MDTYPE valid metadata references

- *LABEL:* This element provides for structured presentation mechanisms when displaying METS documents

- ***Administrative Metadata <amdSec>***

  This metadata section administers the files of which the digital object is comprised of, in addition to descriptions and information related to the creation of the digital object. The metadata are subdivided into four categories summarized below:

  - Technical Metadata <techMD> elements contain information on characteristics of file creation, format and usage whilst sharing the content model common to the <*dmdSec*> section.
  - Intellectual Property Rights <rightsMD> contains license and copyrights as well as information relating to intellectual property rights also aligning the content model to that of <*dmdSec*>
  - Source Metadata <sourceMD>This element also respects the <dmdSec> content model in addition providing descriptions of and administering the analog source objects from which the digital objects are derived.
  - Digital Provenance <digiprovMD>this administrative metadata element contains information on the relationships between sources and destinations

of files of digitized artefacts including transformations and migrations. The element also respects the <*dmdSec*> content model.

- ***File Section <fileSec>***

  This section is more of a listing of all files whose content is related to a digital object's electronic versions. The subsequent subsidiary file section elements are used to group together related and identify elements in general often including a GROUPID attribute for overlapping base information and being classified as follows:

  - <fileGrp> Illustrates the set of files of which a single electronic digital object is comprised of. This element finds relevance and importance when processing large numbers of scanned pages as is in the Jonas Cohn Archive and the other archives evaluated in this dissertation.

  - <file> This element embeds files within a file group <*fileGrp*> with a unique ID attribute for referencing within the METS document whilst at the same time identifying embedded structural and administrative metadata using further ID attributes such as ADMID and FILEID

- ***Structural Map <structMap>***

  The navigational and hierarchical structure of a METS document is defined and encoded in this metadata section. The hierarchy is encoded as a nested series of <*div*> elements hosting either one or more pointers which index relevant and corresponding content. The relevant pointers are classified according to what they are pointing at as either

  - METS pointer <*mptr*>

    Identify corresponding content by specifying a separate METS document hence maintaining reasonably sized METS files.

  - File pointer <*fptr*>

    The focus of these elements is on the specification of relevant files containing content specific information. The specification may be of selected sections of files, groups of files or individual files within the files

section <*fileSec*> of the respective METS document with associated ID attributes.

- ***Structural Links <smLink>***

  The section serves to highlight the existence of hyperlinks between structural map items hence creating archive records of websites by keeping records of the hypertext structure. Whilst the characteristic trait of structural links remains the notion of containing only one element, the elements can be modified using XLink syntax making use of the relevant attributes whereby the “to” and “from attributes” are declared as IDREF.

- ***Behavior <behavior>***

  The main purpose of this section of metadata is to relate executable behaviour with the content stored in a METS metadata object. The behavior <*behavior*> is abstractly defined by an interface definition element implemented and run by a module of executable code which is in turn identified by a mechanism element <*mechanism*>.

2.3.3 Metadata Object Description Schema

MODS Bibliographic element sets belong to the family of bibliographic metadata schema developed and managed by the Network Development and MARC standards office as guidelines for encoding and accessing descriptions of resources in XML format [LC09]. The main idea is to provide for access to resources by structuring their management and discovery as records, in addition to promoting structured management and interchange of the *“encoded descriptions”* i.e. the metadata. MODS schema is considered more a resource description format as opposed to a library standard and hence meant to complement other bibliographic standard schema and a summary of its characteristics includes:

- resource description in XML syntax
- representation of simplified MARC records in XML
- data preparation for metadata harvesting

- more user oriented in comparison to MARC
- element set richer and more user friendly in comparison to Dublin Core and MARC
- ID attributes facilitate element level linking
- facilitates the linking of records with their electronic resources
- inherits MARC semantics
- compliments the METS schema

A MODS document is composed of a root element classified either as *mods* and *modsCollection* as well as at least one element from the set of optional top level elements and attributes. The MODS element order is however not sequential except for sub-elements which can be declared “ordered” and must occur in the given sequence. The MODS root and top –level elements are outlined below:

- **Elements**
  - ***Root elements***
    - *modsCollection*
      a collection of records containing MODS sub elements and the respective attributes
    - *mods*
      an individual MODS record composed of top level elements and attributes
  - ***Top level elements***
    The set of twenty top-level elements described below form the nucleus of the MODS description language. These elements defined as MODS sub elements expand upon the root elements to further describe collections or individual records encoded using the MODS Schema. The characteristics feature of these sub-elements is that they do not specify individual artifact types instead they make use of the *typeOfResource* sub element to characterize the artifact being encoded. The advantage is therefore a straight forward compact vocabulary easy to learn and to implement. As such the top level element vocabulary overlaps with vocabularies of several more artifact specific description languages and hence a candidate for

consolidation in the heterogeneous metadata creation scenario. The list of top level elements includes:

Metadata Object Description Schema – Top Level Elements	
• titleInfo • name • typeOfResource • genre • originInfo • language • physicalDescription • abstract • tableOfContents • targetAudience	• note • subject • classification • relatedItem • identifier • location • accessCondition • part • extension • recordInfo

Table 2.5 MODS Top Level Elements

MODS Mapping and Integration

The Metadata Object Description Schema may be used in conjunction with other Metadata Schema either in complementary form as in the METS/MODS combination implemented for the Jonas Cohn Diaries or for bibliographic mappings for example to MARC [MA04]. As such the MODS semantics are oriented along the MARC 21 Format for Bibliographic Data [MA04] and hence conform to the Open Archival Information System reference model [BB02] [NH09] to be illustrated in chapter 3.

In the METS/MODS scenario the object descriptions are nested within a METS XML fragment assuming the bibliographic record description role whilst METS takes care of the structural linkage to digital versions of the record object. Whilst the MODS schema uses language based tags and a greater number of its elements have MARC 21 equivalents it does not *"assume a display order"* instead using a stylesheet to control the order of record display. However, the MODS schema is open to any cataloguing rules even those outside MARC21 provided each record has its own unique <*identifier*> encoded in <*recordInfo*> or in <*recordIdentifier*>. Whereas identifier elements are unique, top level elements may be repeated and in general the schema does not specify mandatory elements despite MODS documents requiring at least one element.

Element	Element Requirement Level	Subelement(s)/Attributes Required if element used	Subelement(s)/Attributes recommended or recommended if applicable	Repeatable	Content Controlled
<titleInfo> (page 14)	Required	<title>	- type attribute - authority attribute - <subTitle> - <partName> - <partNumber> - <nonSort>	Yes	Recommended authority attribute limits content
<name> (page 18)	Required if applicable	<namePart>	- type attribute - authority attribute - <role><roleTerm>	Yes	Recommended authority attribute limits content
<typeOf Resource> (page 24)	Required	None	- Collection attribute - manuscript attribute	Yes	Yes Recommended authority attribute limits content
<genre> (page 28)	Recommended	authority attribute	N/A	Yes	Recommended authority attribute limits content
<originInfo> (page 30)	Required	- <placeTerm> and type attribute when <place> used - authority attribute when <placeTerm type = “code”>	- <publisher> - encoding attribute for date - point attribute for date - qualifier attribute for date - <edition>	Yes	Recommended authority and encoding attribute limits content

Table 2.6 MODS Summary of Requirements [ME10]

METS/MODS in the Jonas Cohn Archive

In the digitization process of the Jonas Cohn Archive METS/MODS description schema were used to present an ordered digital manuscript version of Jonas Cohn's research and travel diaries. With the internet as the chosen presentation medium, the presentation markup was implemented using the DFG-Viewer [DV10].

```
<mets:mets xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
    xmlns:xlink=http://www.w3.org/1999/xlink  xlmns:mets=http://www.loc.gov/METS/
xsi:schemaLocation=http://www.loc.gov/METS/ http://www.loc.gov/mets/mets.xsd>
  <mets:dmdSec ID="md92018">
     <mets:mWrap MIMETYPE="text/xml" MDTYPE="MODS">
       <mets:xmlData>
            <mods                    xmlns=http://www.loc.gov/mods/v3                    version="3.0"
xsischemaLocation=http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-0.xsd>
             <titleInfo>
                 <title> Pandocheion heteron (deuteron) </title>
            </titleInfo>
            <name>
                <displayForm>Cohn, Jonas </displayForm>
            </name>
          <originInfo>
          <place>
               <placeTerm type="text"> Birmingham </placeTerm>
          </place>
               <dateIssued> 27.01.1946</dateIssued>
          </originInfo>
       </mods>
   <mets:xmlData>
 </mets:mdWrap>
</mets:dmdSec>
```

Fig. 2.19 METS/MODS for the DFG Viewer

2.3.4 Machine Readable Cataloging MARC

The machine readable cataloging format is a standardized bibliographic resource description format developed in the 70's enabling computer based interchange of bibliographic catalog data elements. MARC elements form the basis of modern cataloging formats summarized in 21 formats for *"the representation and communication of bibliographic and related information in machine readable form"* [LC96].The library of congress outlines MARCXML describing it as a framework for implementing MARC in an XML environment. This framework is illustrated by the MARCXML Architecture [MA04] illustrated in Fig. 2.20 below:

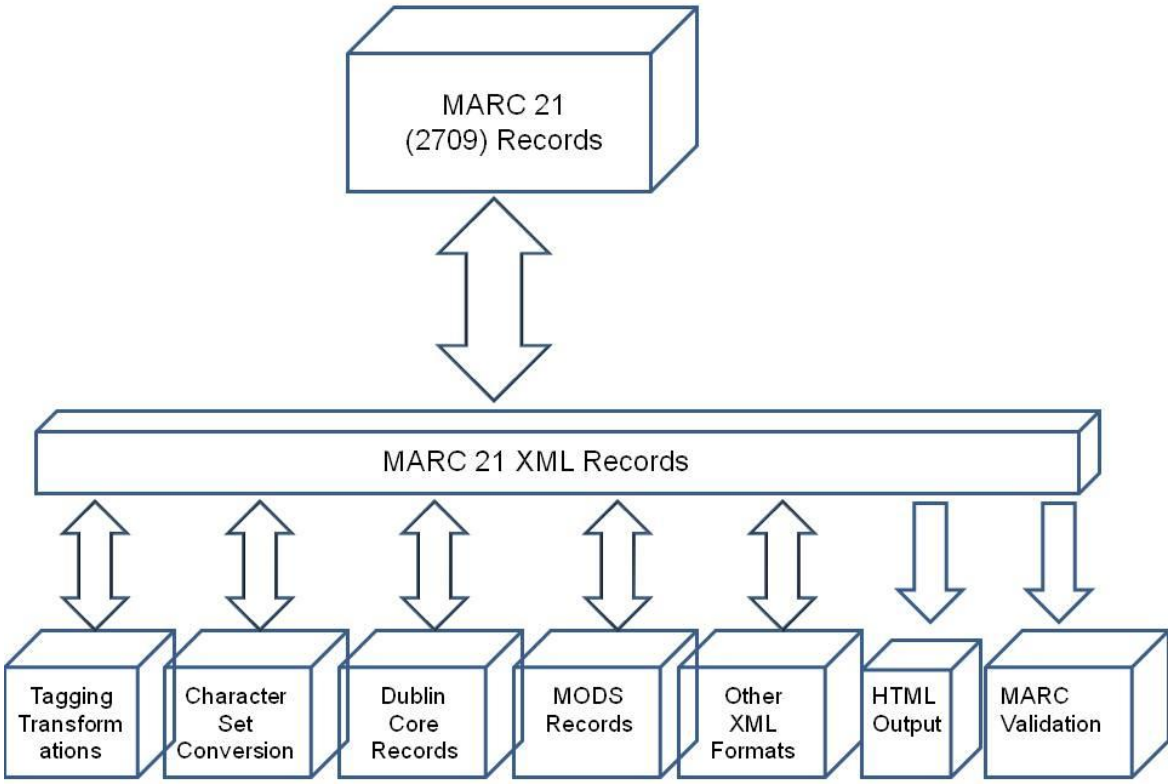


Fig. 2.20 MARCXML Architecture [MA04]

MARCXML Architecture

As a transformation framework, the MARCXML Architecture serves the transition of existing and future metadata encodings from their current formats into the XML domain and the associated standard XML schema in the domain. The architecture is structured in accordance with role as a transformation instrument hence it encompasses components similar to the elements of an OAIS lifecycle summarized below as follows [MA04]:

- MARCXML Design considerations

The design considerations of a component framework determines the component requirements and with respect to the framework purpose. In the case of the MARCXML the framework purpose may be decomposed into eight subsections relating MARC to XML. These subsections map the overall framework requirements to the original bibliographic environment and purposes and are summarized by the Library of Congress as follows [MA04]:

- Simple and Flexible MARC XML Schema

  MARCXML resembles a simple XML Schema containing MARC data whose role is to assume a mediating function between MARC records and other metadata standards. Hence MARCXML resembles a *"bus"* which retains the semantics of MARC21 to reflect MARC in an XML environment. In this context all fields are treated as elements with tags and indicators as attributes, subfield as sub-elements and control fields as data strings.

- Lossless conversion MARC to XML

  Optimal data conversion into the XML environment and back to MARC whilst at the same time providing for non-essential data entries such as MARC structural elements and XML leader data positions

- Roundtripability from XML back to MARC

  Enabling lossless data transitions and record recreation between the different MARC versions via XML transformations (XSLT)

- Data Presentation

  Once MARC data is available in XML, the presentation can be easily achieved by transformations into the appropriate markup.

- MARC Editing

  XML Transformations (XSLT) for updating MARC records by adding, deleting or complementing the record's fields.

- Data Conversion

  XML transformations (XSLT) of data for conversion purposes

- Validation of MARC data

The MARCXML framework validates records at levels with using external software. The validation levels being:

  - MARC XML Schema validation
  - MARC21 field and subfield tagging validation
  - MARC record content validation

- Extensibility

  Implementing MARC in XML enables an open interoperable environment which in turn allows consumer and producer participation in the archiving process where necessary using their archival means.

- MARCXML Architecture

  This architecture summarizes the MARCXML infrastructural domain defining the participating elements as part of an n-tier architecture. The conversion tools are hosted in the higher level tier and tasked with the lossless data conversions between MARC21 and MARCXML. The middle tier resembles a MARC XML bus constituting the *"simple and flexible"* [MA04] MARC XML Schema. This schema is the platform for accommodating MARC consumers and for data transportation. Finally the consumer level defining the three MARC consumer's categories

- transformation
- presentation
- analysis

Transformation referring to the conversion to and from other metadata formats whilst transformation into mark-up or for display purposes and the production of analytical output respectively refer to the presentation and analysis.

- MARCXML Uses and Features

  Using MARC as a structured data scheme has several advantages for any digital archiving project. In addition to the flexibility of the MARC XML framework the constant maintenance and support by the developing institution, The Library of Congress [MA04] is almost certain. As such the summary of the uses of MARCXML includes:

  - representing bibliographic records in MARC and XML compatible forms whilst in compliance with OAI and its data harvesting infrastructure
  - extending the METS schema whilst packaging the metadata with an electronic resource

Furthermore, MARCXML resembles a platform supporting the processing of data encoded in all MARC formats whilst allowing plug and play to custom user solutions.

In addition to the structural outline of the architecture illustrated above the MARCXML standard is maintenance includes support and transformation sheets for conversions to other standard XML schema namely:

- MODS Conversions
  - MARCXML to MODS
  - MODS to MARCXML

- Dublin Core Conversions
    - MARCXML to RDF
    - MARCXML to OAI
    - Dublin Core to MARCXML
- OAI MARC Conversions
    - OAIMARC to MARCXML

2.3.5 Encoded Archival Description

The latest version of the EAD was developed and presented by the American Society of Archivists in 2002 as a data structure for encoding finding aids and providing user access to archive resources from a multitude of institutions. The design principles focus on structuring archival descriptions for data communication and resource sharing with the help of universal tag elements however not at the presentation level. The 146 description elements are organized within a tag library generally conforming to the Open Archival Information System reference model.

The figure 2.21 below describes the EAD tag conventions illustrating the structure of an EAD document in addition to the three basic EAD group elements <eadgrp><archdescgrp><dscgrp>. An EAD document is generally hierarchically structured commencing with the tagged name at the top followed by the full element name and meaning. The general structure can be listed as follows:

- Tag name
- Element name
- Description
- May contain
- May contain within
- Attributes

EAD Tag Elements

In contrast to the other metadata structures presented in this work, EAD tag elements are listed as library and therefore not classified according to the object focused description sets a few examples of the tags include:

- <*filedesc*> - File Description
- <*fileplan*> - File Plan
- <*persname*> - Personal Name
- <*processinfo*> - Processing Information,
- <linkgrp> - Linking Group

EAD Tag Attributes

Whereas the tag elements are not organized in element classes, the tag attributes are clearly classified as either being:

- Linking Attributes

  Apply to elements used for linking and may include hyperlink descriptions, resource identifications and entity references

- Display Attributes

  Apply to structural presentation elements and are used to format presentation structures such as tables and columns

- General Attributes

  Apply to all EAD elements reflecting the named characteristics of the element whilst being hosted within the brackets of the element tags.

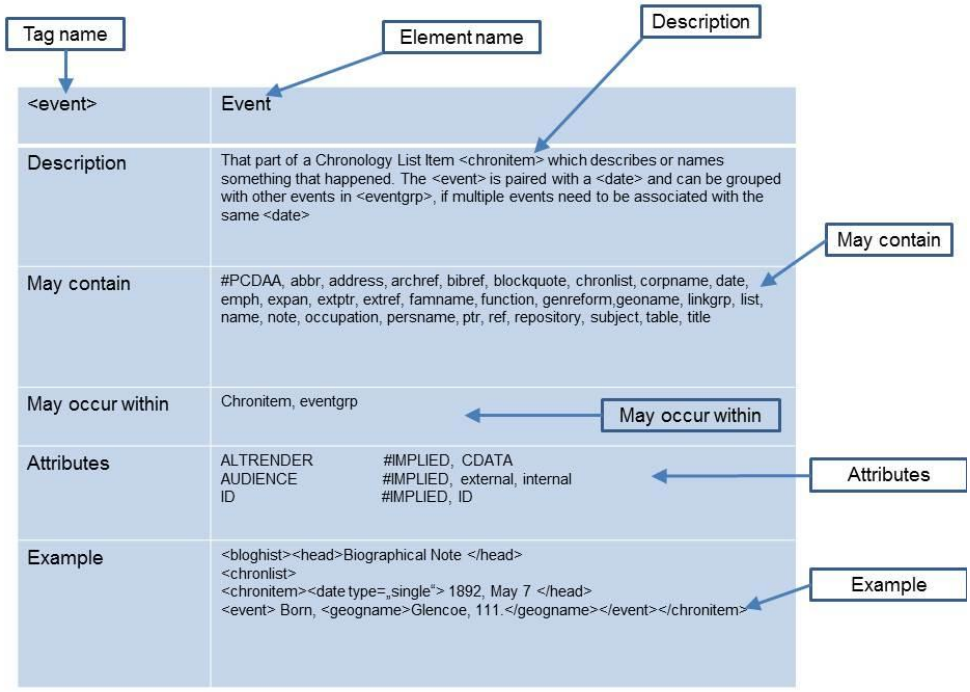


Fig.2.21 EAD Tag Conventions [EAD09]

2.3.6 Text Encoding Initiative

Text Encoding Initiative TEI is widely accepted as the XML standard for text editions and machine readable texts illustrated in digital form. TEI finds broad application in digital epigraphy, transcriptions and scholarly digital editions marking up a wide range of corpora on the basis of an extensive element definition set summarized by the TEI guidelines. The current P5 version consists of 20 tag element modules supporting the compulsory TEI infrastructural elements in their descriptions of digitized documents. The element tag sets include the following modules:

- Verse
- Performance texts
- Speech transcriptions
- Manuscript descriptions
- Dictionaries
- Bibliographic critical apparatus
- Names, Dates, People, and Places
- Tables, Formula, Graphics and Notated Music
- Linking, Segmentation, and Alignment
- Graphs, Networks, and Trees
- Language Corpora
- Certainty, Precision, and Responsibility

The list above shows just a cross section of the different elements defined in TEI and hence the wide application field of this description language. Interesting to note is the overlap between content and format as shown by the existence of categories for Tables and Graphics as well as those for transcriptions and manuscript descriptions. In other words TEI descriptions appear to overcome the separation of content and format implementing both within the same vocabulary space. The figure below shows the TEI modules and identifiers in association with their application areas

Module name	Formal public identifier	Where defined
analysis	Analysis and Interpretation	17. Simple Analytic Mechanisms
certainty	Certainty and Uncertainty	21. Certainty, Precision, and Responsibility
core	Common Core	3. Elements Available in All TEI documents
corpus	Metadata for Language Corpora	15. Language Corpora
dictionaries	Print Dictionaries	9. Dictionaries
drama	Performance Texts	7. Performance Texts
figures	Tables, Formulae, Figures	14. Tables, Formulae, Graphics and Notated Music

Fig. 2.22 TEI Element modules [TEI09]

3 System Requirements Analysis

An analysis of the requirements of a system is prerequisite to the system's specification and the subsequent results constitute the basis upon which the system is implemented. However, the definition of the term *system* is necessary for limiting the boundaries of the analysis and the respective environment. In the German Industrial Norm (DIN 19226) a system is defined within the framework of a classification of unit sets in direct or indirect relation to each other with the set of relations between the unit sets being referred to as the *structure* of the system in question [SD71]. McCormick [EM79] incorporates humans and interactions in his definition of a system supporting the definition of an information system and its inclusion of the former together with information technology to result in a composite environment aiming to solve given tasks. With this combination representing an *"application landscape"*, working to produce a specific target product the individual tasks of the participating units become the subject of the system requirements culminating into a set of user tasks accompanied by the system implementation.

An investigation of these requirements subscribes to De Marco's definition of the *"study of a problem, prior to taking some action"* referred to in general as an analysis whilst *"the study of some business area or application, usually leading to the specification of a new system"* as analysis specific to system development [TD79]. The latter relates a system to its requirements and specifications and its investigation resembles a system analysis which is composed of an analysis of the application landscape and its requirements in this case the framework and tasks allocated to the user illustrated by way of a user task model. This chapter will study the problem *"metadata creation for digital archives"* prior to a specification and implementation of a framework for creating such metadata i.e. *"the new metadata creation system"*. Whereas the specification of the system is reflected by the use case, the content and user task specifications are respectively illustrated by the data model and the user task models dealt with in the succeeding subsections of this chapter. The latter is further associated with the formative evaluation also described in the coming subsections but elaborated in detail in chapter 6.

3.1 Framework Analysis

The primary concern of structurally analysing a metadata creation framework system lays within the identification and classification processes [AG05] of the information to be collected i.e. the *functional requirements*. These model-based processes resemble an analysis and specification of the interactive digital archive software tackling semantic aspects of the development [P99]. On the other hand digital objects in object-oriented models are comprised of structural and behavioural characteristics respectively describing attributes and operations. Paternò [P99] defines such digital objects as *"entities manipulated to perform tasks"* either as *perceivable* objects or *application* objects. The human-computer interactive nature of data capture archive software also steers the focus towards user-oriented aspects in form of tasks and their relationships for individual user types. Indeed these tasks and their relationships encompass and capture the user interface requirements of the archive system being developed. In this dissertation the system is modelled using the unified modelling language (UML) with the respective functional and task analysis requirements aspects being catered to within the framework of the following tools:

- *Use cases*

  models of the system functionality and its environment in support of the archival processes. Muller et al. [RM99] summarize the resulting effects of use cases on the system design as:

  - representations of atomic transactions through the framework relevant to the architecture accommodating system processes.
  - illustrations of internal and external data manipulation in addition to the structural organization of the data elements concerned.
  - providing the principles for the validation of the digital archive framework
- *Interaction diagrams*

  object oriented diagrammatic illustration of the sequence in which requests between objects are executed. Useful for understanding queries and for building indexes.

- *Activity diagrams*

  Illustrate the high-level view and flow of a process in addition to the business, in this case archival operations.

- *State charts*

  Capture the dynamics of the interacting objects within the archive system

- *Class diagrams*

  Depict the underlying structure of the system as logical system models

- *Component diagrams*

  Include applications and interfaces used to access a respective database in addition to illustrating the database management structure.

3.1.1 Metadata Creation Requirements

Understanding the functional requirements of the metadata creation framework begins with an understanding of the framework's users, their information requirements and the activities involved. Rosenthal et al. [RR05] identified the goal for digital preservation systems as being *"that the information it contains remains accessible to users over a long period of time"* and hence eliciting the requirements criteria categorized as replication, migration, transparency, diversity and economy. At the same time Rosenthal et al. [RR05] acknowledge that state of the art digital preservation activities have a focus on metadata and standards as opposed to the actual content. Wright [RW06] goes on to define the role of an archivist in the digital world as that of managing the content i.e. *"concentrating on the metadata (catalogue and other finding aids rights data) required to manage the content - and defining the requirements for* ***storage services****"*. Whilst extracting metadata from content proves to be efficient, hand-generating form metadata and assistance in form migration shows remains the key activity despite efficiency and economic deficiency. In return, Hodge et al. [GH00] summarizes the *"best practices"* for digital archiving in terms of metadata creation as being based upon:

- metadata data types
- standards and interoperability
- resources and content rules

- application level and purpose

As such these metadata creation characteristics form the foundation for a best practice metadata creation at the *"object creation stage"* or in cases where complimentary metadata are created in subsequent stages *"with metadata provided at creation being augmented by additional elements"* [GH00].

Furthermore, looking at documents and image collections where hand-generated metadata creation is *"not sufficiently incorporated into the tools for the creation of these objects"* incorporating these requirements as part of the creation framework simplifies the novel collection process. Hence metadata creation requirements based on these characteristics resemble functional encoding facilities assisting the management of content as categorized by Rosenthal et al. [RR05]. In so doing they specify the archive service and functional storage requirements and challenges in line with Wright's analysis of digital archiving requirements as [RW06]

- ***persistence*** : *"the ability to get content out of storage"*
- ***currency*** : the ability to utilize stored content
- ***preservation actions***: mediated interoperability
- ***obsolescence***: the characteristic of being replaced in light of technical advancement

In other words a metadata creation framework facilitates the hand generated record collection and object description activities of a digital archivist managing digital content at the object creation and augmentation stages. ***Persistance*** requirements dictate data structuring, in our case based on standard XML schema resulting in turn in ***currency*** requirements being fulfilled however associated with necessary data entry interfaces.

The state of the art processes involve text markup and object linking within a client server web-based architecture enabling an encoding and archivist suited markup environment. The novel schema independent ***preservation actions*** require semantic schema mediation within the framework accommodating data abstraction and modelling in addition to an object schema mapping. The mediation process tackles XML and schema – based ***obsolescence*** based requirements and future compatibility requirements.

The metadata creation therefore requires an automated mediated meta-tagging interface mediating between the descriptive data, the content and the respective digital object. Whilst the tagging facility serves hand-generation of collection and object description data, the mediator maps the respective data to the necessary XML schema. With the archivist assuming the role of a content manager, the encoding context of descriptive data creation process should remain encapsulated in the application system resembling hand generation. Heterogeneity requirements result in the need for creation instances generating described metadata in one of the respectively requested schema.

The requirements dealt with above consider the overall necessities associated with the creation of structured heterogeneous metadata in a digital archive environment accommodating long term preservation. However, one of the most important requirements for preservable metadata lies in obsolescence through the data abstraction and meta-language independent encoding. Having the creation framework deal with this phenomenon implies the need for technical requirements for translating the abstract data elements into structured meta-documents. These requirements define the metadata management character of the framework and refine the above mentioned requirements incorporating the relevant meta-language vocabularies. The vocabularies reflect the short term encoding needs and are represented by associated user tasks which in turn have to be taken into consideration when analyzing the creation aspects. These needs can be illustrated as java binding instantiations of the object classes modelled within the framework and expanded upon in the coming chapter. A summary of these instantiations illustrates their relation and translation into Wright's [WR06] digital archiving requirements and the underlying criteria for best practice digital archiving according to Hodge [GH00]. Accordingly metadata creation takes place at one or more of the following digitization stages object creation, augmentation and object mapping stages. Since the objects are modelled and implemented as java classes, the framework system requires class instances for creating XML metadata using business data related to the digital object in question. These metadata creation instances can be summarized as follows:

- **Object Creation Instance**

  Description metadata at data entry level in line with task and module oriented schema vocabulary dictated by object type and purpose as well as its relations to other objects and attributes. This creation stage generally resembles a population of java business objects meant to be mapped to respective XML schema for preservation and data interchange purposes.

- **Augmentation Instance**

  At this instance, metadata have been modelled and implemented as java classes and now a modification of the context and content of existing java classes and their respective business objects takes place hence a repopulation of existing business objects. The respective XML marshalling complements the creation process for preservation and interchange purposes.

- **Data Management Instance**

  Invoking this instantiation associates business objects unmarshalled from a base schema into java being marshalled back to the relevant second schema and catering for semantic heterogeneity. The creation requirements include the relevant XSD schema file and the set of object instances for conversion. Data interoperability and data interchange follow the modularized digital objects and their classes being marshalled into the relevant meta-language schema for standardized interchange.

- XML Schema

  The corresponding data structures associated with business data objects created in the metadata framework require a specification of the XML structure and data elements. For the purposes of this dissertation, specification will be via XML Schema files in XSD format. This incorporates the requirements of the binding scheme for XML <=> Java <=> XML enabling the transformation of java business objects into structured XML documents for digitization and preservation purposes. The binding scheme modularizes and mediates between the java classes and the XML elements of the relevant digital archive standard. Whilst a base schema for the initial unmarshalling of the XML schema into java object classes is required, instantiations of these classes into another XML is sufficient for further java to XML mappings.

This corresponds to a low level semantic mediation between heterogeneous XML Schema with a homogeneous semantic content modularized via the digital java business objects and their classes. XML Schema allow type specification in addition to pattern matching and extension [BM02] hence accommodate data types and constraint metadata models in line with the “best practices” [GH00].

Creation Requirements Summary

- **Abstract creation of structured XML documents for the management of digital content**
  - The abstraction should support hand-generated metadata creation enabling without familiarization any knowledge or familiarization with a meta-language
  - The creation should cater for the structuring and generation of the metadata in structured XML from the abstract digital objects
  - Heterogeneous task oriented structuring and encoding should encompassed Object Marshalling
  - This system should support the specification of digital object types and metadata elements and attributes as integers, strings or characters.
  - Illustrate the text image nature of the archive material
- **Content and context elements must be explicit and business oriented avoiding the duplication of structural elements this means**
  - a closed structural vocabulary i.e. the business objects – constraint data model
  - relate digitized object to archive record
  - harmonise metadata elements with the business objects
  - binding the class object and meta-structure
- **Map abstract digital objects to standard XML schema whilst mediating between standards. This mapping should**
  - fulfil respective digitization standards
  - navigate metadata between XML schema
  - marshall digital business objects to XML structure
  - unmarshall objects to java for instantiation

Table 3.1: Summary of Metadata Creation Requirements

In software engineering terms the metadata creation requirements discussed in the preceding subchapter represent a summary of part of the concept [RW06] and conceptual model of the digital archiving business. Embracing this conceptual model [NM01] in the unified modelling language associates the concept of digital archiving with the business use case modelling and business object modelling necessary for modelling the framework system. Identifying the digital archivist as business use case actor and the researcher as archive user results in the metadata creation use case illustrated in fig. 3.1 for the abstract creation concept. Whilst unmarshalling is a part of the business use case, its role remains that of object marshalling defining of the structural composition of the metadata based on the java objects. The respective descriptive data in XML constitute a part of the framework business use case together with the mediation concept for semantic heterogeneity and are dealt with in the succeeding section on framework requirements.

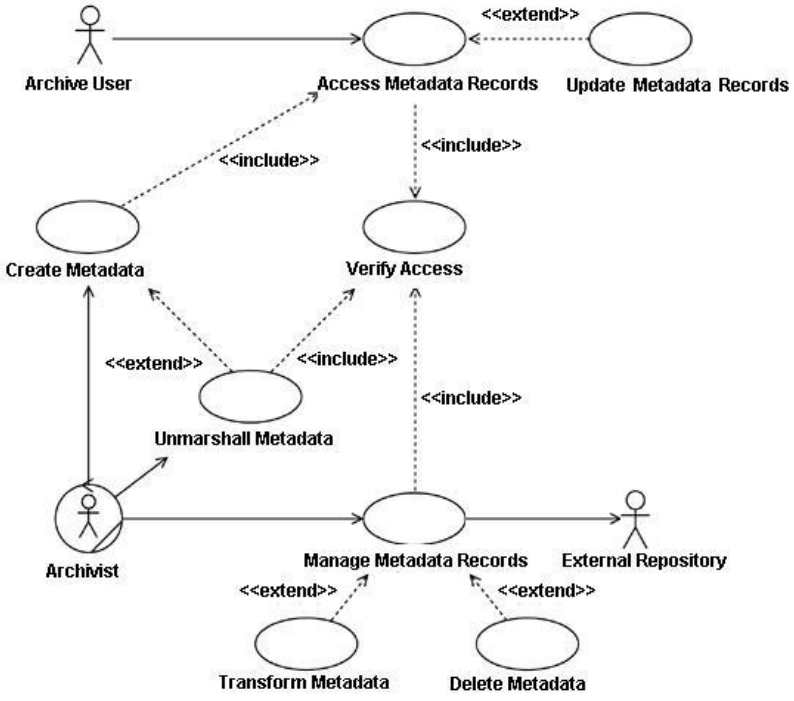


Fig. 3.1: Metadata Creation Use Case

3.1.2 Mediation Requirements

The notion of an interface supported framework mediating between archivists and their digital archiving and interoperability requirements illustrate the state of the art digital preservation business. In this dissertation, this notion constitutes the idea of schema and hence metadata heterogeneity whereby the framework mediates between digital business objects and their metadata in XML. The supporting concept enabling such mediation is the Java XML Object Binding which will be elaborated upon in chapter 4. The mediation concept involves invoking object mapping and metadata transformations as well as the generation of the descriptive preservation data as part of the framework business. As such the conceptual model illustrated as a business use case below resembles a closed system initiated externally by the system user, in this case the digital archivist. The conceptual requirements of such a system summarized in the table below:

Mediation Requirements

• **Conversion of digital objects into XML metadata elements**
populated abstract digital objects hand-generated by the archivist need to be preserved in the XML meta-language. The business of data binding is carried out by a binding framework and this framework requires:
  • ***binding schema*** [BM02]
 	
• **Modularizing XML schema elements and attributes**
heterogeneous XML elements containing the same business data need to be normalized by modularizing the contents and context in the business objects. This can be done with the help of metadata constraint models and hence the need for the definition of:
  • ***constraints*** [BM02]
  • ***content specification***
  nesting content elements consisting of “other element references, choices and sequences” [BM02] for normative references
 	
• **Semantic equivalence by mapping business objects to heterogeneous XML Schema**
The structuring of the same via different formats requires **process loops** [BM02]

Table 3.2 Mediation Requirements

The focus of the framework illustrated here by the mediation requirements is on the creation of the XML structured documents populated by business data as described by the framework classes. The binding schema and the constraint model resemble the backbone of the XML transformation of the metadata stored as class based business objects. The description element modules and the content specification regulate the inter-schema transformations and crosswalks whilst at the same time stressing the business orientation in opposition to the schema orientation. As a result the conceptual requirements summarized above are illustrated by the use case below whilst the system requirements of the semantic heterogeneity in the form of process loops are illustrated in the figure 3.2 below.

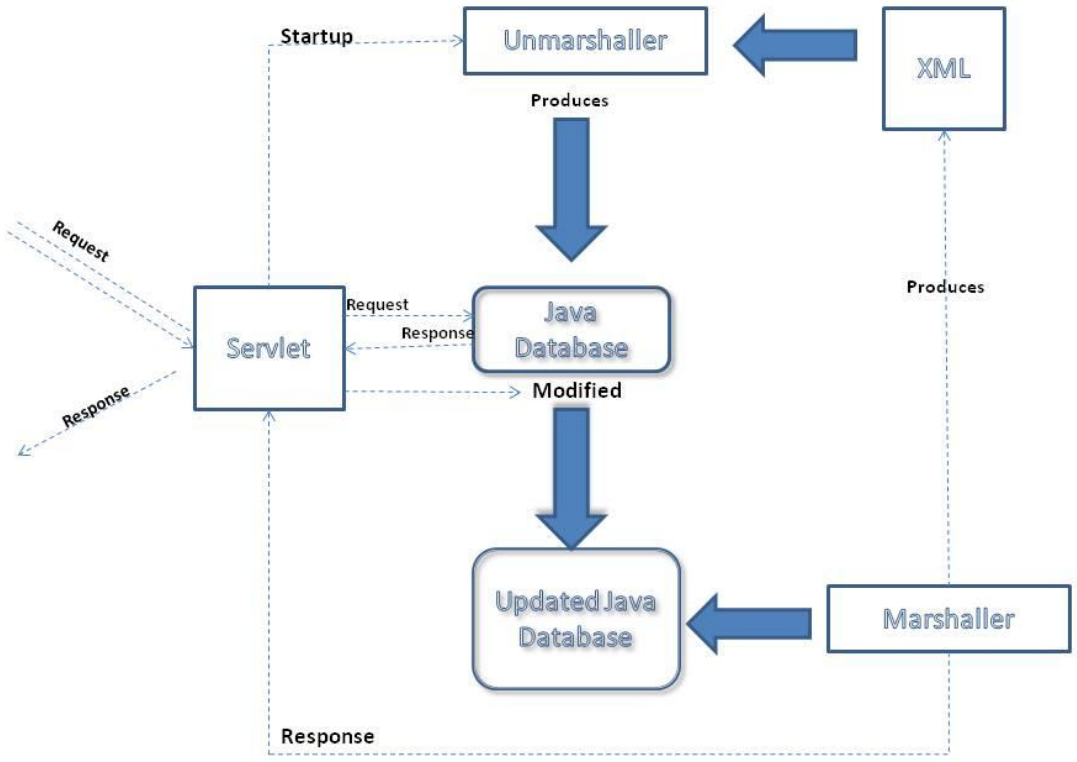


Fig. 3.2 Process Loops in Java XML Binding [BM02]

It should be noted that the process loops mentioned above represent a requirement of the mediation framework. Thereby in accordance with the definition of process loops, the output of associated processes also serves as process input i.e. the engineering concept

of feedback. The respective processes for this metadata framework are described in chapter 3.2 with process loops consisting of:

- **Unmarshalling process**
  - Class generation process
  - Validation
- **Marshalling process**
  - Element generation process
  - Validation

Both processes and their sub-processes are part and parcel of the data binding framework with the respective data and constraint models for the metadata semantics.

Preservation Requirements

The business concept of preservation in digital archives resembles the concept illustrated by the metadata creation framework developed for this dissertation. Hereby elicit interfaces and processes provide the means and ways of holding and enabling long-term access to digital objects. The preservation requirements of the framework itself are then reduced to the attributes and accessories necessary to implement the standard operations of the framework. These attributes include the selection of relevant preservation metadata standards and the provision of abstract interoperability platform and in so doing dealing with the problem of obsolescence and the need for replication. Fig 3.3 summarizes the functionality required of preservation metadata with reference to digital content description within the context of long-term accessibility. This preservation effort [DE10] summarizes the functionality in function types aggregated against structural and content variation in preservation spaces. The figure leverages agnostic metadata ignorant of content and organization, descriptive metadata acknowledging the structural artifact and the content specific metadata identifying underlying digital object structures text, audio, video and facsimile image. In other words; managing the content and *"concentrating on metadata required to manage the content"* [RW06] whilst defining storage requirements with respect to replication, migration, transparency, diversity and economy [RR05] constitute the preservation activities. These activities can then be summarized as follows:

Preservation Needs
Management policies and infrastructure for providing storage and long term access to digital records and their content. With the framework being part of preservation infrastructure, the refined requirements of the infrastructure can be summarized as: • ***Preservation metadata*** Descriptive, administrative, structural and technical metadata documenting ownership, digitization, state of the art, and access rights specifying preservation goals which Dappert et al. [DE10] summarize as: o Digital control is within the physical control of the repository o Digital content can be uniquely and persistently be identified and retrieved in the future o All information is available so that digital content can be understood by its designated user community o Significant characters of the digital assets are preserved. o Digital objects remain whole and unimpaired and that it is clear how all the parts relate to each other • ***Replication*** Grid or multiple storage activities or participation in network storage activities • ***Redundant systems*** Multiple documents of the same structured metadata collections with respective consistency validation mechanisms avoiding *"single point of failure modes".* • ***Automation*** Reduces human cost of procedural documentation of data ingestion, storage and distribution. • ***Multivendor software and hardware facilities*** Heterogeneous software facilities for software and format independence • ***Interoperability*** Supports the replication and redundancy activities with data import and export for and from network storage activities

Table 3.3 Preservation Needs

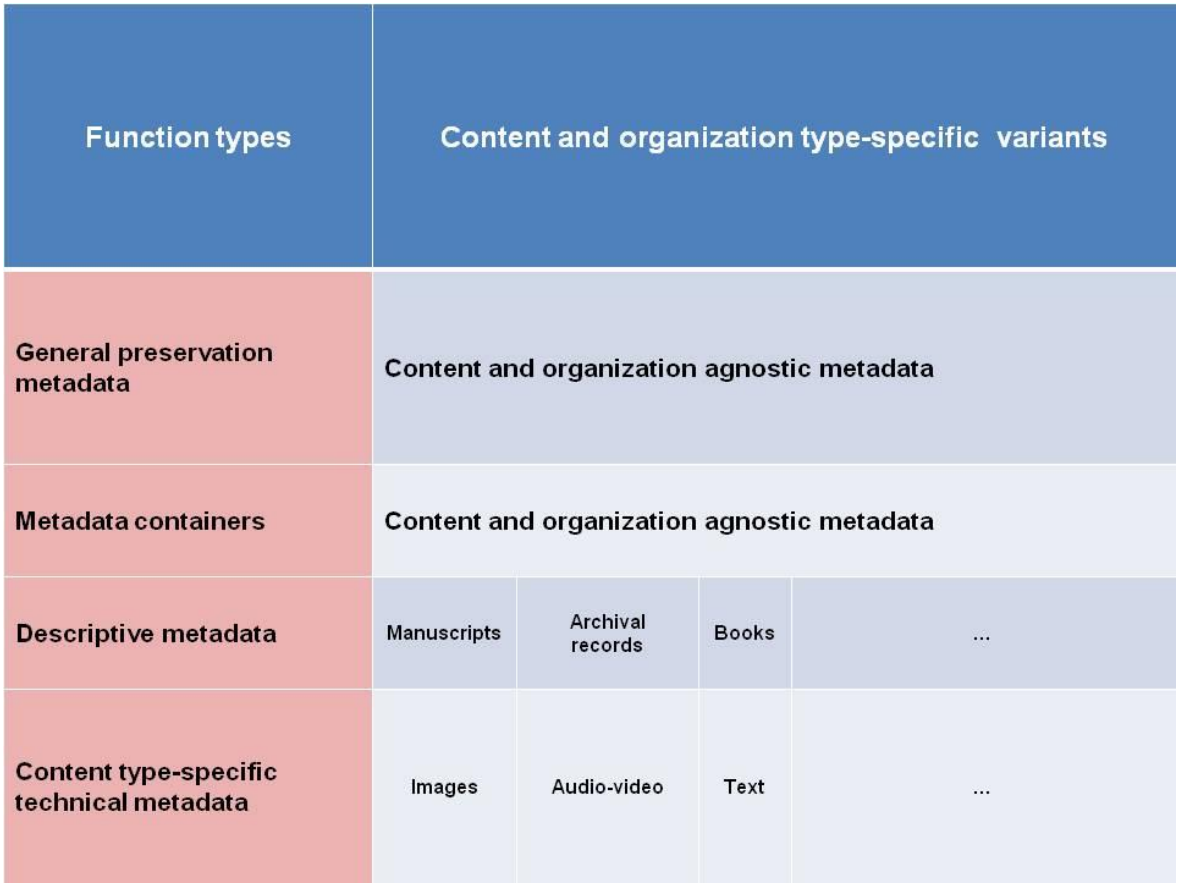


Fig. 3.3 Digital Preservation Metadata Efforts [DE10]

Interoperability

The requirements for data interchange and interoperability lie within the mediation effort of the metadata creation framework. Interoperability forms the basis of meta-language based preservation activities in general and for participation in grid storage and catalogue networks. This basis relies on crosswalking XML metadata to suit the heterogeneous formats of the semantic descriptions of the archive metadata. Common state-of-the-art formats include Metadata Encoding and Transmission Standard (METS) and its associated Metadata Object Description Standard (MODS), Encoded Archival Description (EAD), MAB and MARCXML and of late Resource Description Framework (RDF) for semantic linked data. A selection of these formats has been elaborated upon in chapter 2.3 on structured data schemes. As such the mediation effort and therefore interoperability lies within a modularization of the different metadata elements sharing the same business descriptions of the digital archive. The modularization needs can be summarized as the business data model needs of the archive. The data model needs are defined according to the constraint models and described in chapter 3.1 and are

summarized by the Rules and Recommendations for Archives (RNA) [WK10] elaborated upon in the subsequent chapter on Data Modeling.

3.1.3 Archive Framework Task Analysis

User-oriented software development activities rely on task analysis and task models for investigating user interactions and sequences requirements underlying a proposed user interface [P99]. The model identifies interaction elements together with their relationships whose implementation by means of a user interface facilitates communication with the application. Furthermore, the interactions may be distinguished from a user point of view as either being content or process oriented [BB07] thereby determining the structure of the interface. As a result this structure focuses on either tasks, content or user roles with these elements dominating the design of the interaction processes. At the same time state modifications and event-driven interaction elements can be analysed using task allocation based hierarchical trees in a task model. The associated tasks for such a model specifying a digital archiving framework are related to the role of the system user, in this case the digital archivist as defined by Wright [RW06]. The role of such a system user focuses on metadata based content management for preservation purposes requiring system ***persistence*** and ***currency*** [RW06] hence an overlap of content and process orientation. Task models for systems illustrating such orientation overlaps are characterized by Bomsdorf [BB07] according to the following three perspectives:

- ***usage oriented-processes***
  representing task and activity procedures from a user perspective
- ***purpose-oriented processes***
  reflecting the purpose and goals of an application from a business point of view and often represented by business processes
- ***system based processes***
  principle aim is the implementation hence specification by internal control in addition to the business logic and the respective business processes

Mediated Archiving Tasks

In the requirements analysis above, we analyzed several different facets of the descriptive set of requirements necessary for creating and structuring heterogeneous XML metadata for interoperable digital archives. We have ascertained the requirements for abstract metadata acquisition, the resulting heterogeneous XML schema encoding and the preservation aspects of the digital archiving process. The analysis reflects upon the mediation process of the digitization framework commencing with the data acquisition before encoding the preservation metadata as the framework output. Acknowledging the process implies the need to analyze the tasks specific to metadata acquisition and mediation process in the context of modularized semantic heterogeneity. As such, these tasks involve interaction tasks between the system user creating, the acquisition modules and database and the resultant XML documents. Paterno et al. [P99] defines the identification of such tasks as follows: *"interaction tasks themselves can be identified as user tasks and application tasks, as in use cases."* As a graphical user interface framework results of the interaction requirements analysis modelled within a task model encompass the basis of the user interface and interaction design. However the purpose of task models is to describe system user interaction activities resulting in the achievement of the user's goal whilst incorporating the foreseen requirements [P99], in our case the metadata creation requirements. As such the tasks represent the activities performed to achieve a goal and an analysis of which is then used to develop abstractions of the task model. Paterno et al. sees the purpose of the analysis as *"to identify what the relevant tasks are"* and summarizes the analysis techniques as follows:

- Interviews or workshops
- Questionnaires
- Observing user in their workplace
- Considering how activities are performed in the current environment
- Considering existing documentation and training methods

The task identification process for digital archiving described in this dissertation is hence related to formative evaluation described in chapter 6 whose purpose includes among others instructional and interface design. This formative evaluation considered current metadata encoding activities and interviews with a sample of archivists maintaining hand-generated scholarly digital archives.

Formative Evaluation Interviews

The main objective of the formative evaluation interviews was to judge the necessity of a metadata encoding framework in general and an associated graphical interface for encoding i.e. the usability requirements in particular, with the results serving as feedback for the system during the development phase. The evaluation objectives can be made out to be:

- eliciting archiving tasks and goals
- eliciting metadata element categories
- determining current preservation activities
- determining whether structured archiving is taking place
- determining which structured data schemes are in use

The formative evaluation dictates an indirect assessment of the framework problem description as the test users' familiarity with the notions of modularization, semantic heterogeneity and data abstraction in the computing sense. This implied that the interviews of the formative evaluation had a focus on the user tasks, the existence of structured data collection mechanisms and the description categories required to implement the overall preservation and archiving goals. As a result the overall evaluation objectives were all in all an assessment of the results of the formative evaluation. In return these results us an insight of the extent of structured archiving activities in scholarly archives providing an empirical foundation necessary to determine the evaluation and hence prove the following concepts:

- usefulness of modularized data elements
- semantic heterogeneity and the relationships between digital entities
- suitability of the concept of mediated abstract metadata creation

Since the framework design and implemented here is a prototypical implementation of the concept of mediated abstract heterogeneous metadata creation, the formative evaluation serves as a proof of concept meant to expose conceptual weaknesses with the help of flexible evaluation methods [LMB02]. The flexible method involves an assessment of the suitable empirical data collection methods from the palette introduced by Paterno and illustrated above. This palette assumes the necessity of technical assistance and hence the simplification of complex and electronic media supported processes and activities. The selected test design is described in general below before an analysis of the evaluation design is allowed to influence the task analysis and modeling process.

Test Design

The formative evaluation of the metadata framework interface involved a heterogeneous sample of eight archiving scholars whose activities included managing archive content. The sample scholars represent four system user categories in accordance with Nielsen's heuristic evaluation test user theory classified according to their purported archiving goals also subject to assessment. Whereas digital archiving represents a common activity for the test users, the intended goals and current processes aimed at achieving those goals do differ. The test users were not expected to have technical background without however excluding those possessing such a background hence reflecting the real as opposed to the theoretical and normalised test case. All sample test users work with intellectual archive material within the framework of academic research in the field of humanities and their record collection activities can be classified according to their contents within the context of:

- scholarly correspondence
- epigraphy
- literary editions
- cataloguing

The test itself resembled carrying out exemplary record collection tasks involving an intellectual product from the test user's own archive material using their normal archiving tools. The tests took place in the user's controlled natural environment within a regulated time frame and were repeated for the same data sets but this time following guided by the

rules and recommendations for archiving RNA [WK10] and Functional Requirements for Bibliographic Records [FR08] where applicable.

The test user was then required to repeat the same activities mentioned above, only this time not only with a data ser from his/her own archive but with also from an archive dataset from another collection perspective and this time using the prototype archiving framework. The purpose of the design was to elicit the test user's normal archiving tasks and activities, whilst assessing the use of standardisations and recommendations supporting information interchange and the concept of open archives whilst investigating the user awareness to the metadata background and encoding basis and aspects of the digital archiving activities.

The results of the test user's experience were reviewed in an interview resembling the questions contained in a questionnaire, the interview being documented photographically. The informal interviews for user testing purposes were mainly as a result of requests expressing preference in being interviewed as opposed to filling out a questionnaire as part of the empirical evaluation. An appraisal of the interviews presented the results of the formative evaluation, categorized according to archiving tasks, archiving standards and framework/tool usability. As a prototype the concept of usability and user experience were united to avoid terminological confusion between heterogeneous test user groups.

Results

The general aim of a formative evaluation with test users is to obtain user feedback with respect to the proposed concept of a framework for metadata creation and usability aspects related to utilizing such a framework. This procedure is commonly referred to as a proof of concept despite the extra aspect of utilizing user feedback as input for the task analysis and application development processes. The development process then represents a set of correlated variables illustrating the tasks interacting within a closed system and resulting in a competent framework application.

The qualitative informal feedback consisted of an interview reviewing semi-structured questions intended for a questionnaire. The tables below summarize the evaluation results of the interviews

Q. What are your archive usage goals?	
Response	**% Users**
online presentation	57%
digital edition	43%
book edition	28%
secondary source	71%
non	0%

Q. Which categories of product entities do you record?	
Response	**% Users**
manuscript	86%
correspondence	29%
lecture	43%
journal	71%
diary	14%
other	29%

Q. Which categories of responsibility entities do you record?	
Response	**% Users**
author	100%
custodian	37%
recipient	25%

Q. Which categories of subject entities do you record?	
Response	**% Users**
text	0%
text images	0%
optical files	29%
born digital sources	71%
sermon	14%

Q. What are your long term preservation measures?	
Response	**% Users**
microfiche	0%
film	0%
optical filing	29%
storage files	71%
born digital	29%
non	14%

Q. Which categorisation techniques do you employ?	
Response	**% Users**
card index	14%
MARC-based	29%
XML-based	29%
Office-based	71%
Non	14%

Q. How often do you participate in archive catalogue networks?	
Response	**% Users**
regularly	29%
average	29%
irregularly	71%
seldom	14%
no idea	14%

Q. Are you familiar with PND Number elements?	
Response	**% Users**
yes entirely	14%
yes partially	57%
not sure	43%

Table 3.4 Formative Evaluation

3.1.4 The Task Model

One of the main aims of developing a framework supporting metadata creation is to enable wider access to and the ability to structure archive metadata. The underlying interest of doing so focuses on simplifying the metadata encoding and its respective processes and therefore requires the implementation of measures accompanying the system design meant to support system – user interactions. The notion of modeling interaction processes from a usage point of view illustrates the background and concept behind the *"logical descriptions of activities*" performed to reach a goal [P99] [JN94] i.e. task modeling. According to Paterno et al. [P99] task models resemble structured methods allowing system designers to manage complex usability factors when designing interactive applications. Nielsen expands upon this definition to include an analysis of how users *"approach the task, their information needs and how they deal with exceptional circumstances or emergencies"* [JN94]. Although task models in general focus on interaction models for usage, domain or system-oriented processes [BB07], the models relevant to this metadata creation framework follow the usability and usage process lead due to their modeling of tools simplifying encoding tasks.

Translating Wright's definition of the archivist's role in the digital world to the interaction world of tasks and goals requires eliciting digital archiving activities before assessing how these can be performed whilst interacting with the framework to reach the digitization goals. This includes incorporating the requirements and perspectives of the actors involved in the digital archiving process.

The results of the formative analysis and the requirements outlined in the preceding subsections belong to this group of factors and will be taken into consideration whilst designing the task model for the framework assisted heterogeneous metadata creation use case. Since tasks and goals and directly related to each other, *"each task can be associated with one goal, that is the goal achieved by performing the task"* [P99]. In other word the classification of the type of task model is subject to the design goals and here Paterno et al. [P99] identifies three main task model categories:

- the system task model
  also referred to as the functional analysis [JN94]describing the structures of system use and functionality whilst assuming task implementation procedures from a system point of view i.e. the functional reasons for the tasks.

- the envisioned task model
  describes proposed system interactions of newly developed system leaving room for further definition.

- the user tasks
  defines user specified task descriptions in light of associated user goals extracting particularly effective users including their strategies and *"workarounds"* [JN94], hence the relatively flexible model structure depending on the users in question. Paterno et al. [P99] view the discrepancies between system and user task models as the source of the bulk of usability challenges. The users' model also serves as a *"source for metaphors for the user interface"* [JN94]

Modelling requirements of a digital archiving framework presents software development with the unique task of integrating characteristics of content and process oriented interactions into a unified task model. The table [BB07] below summarizes the characteristics of the respective interaction types identifying activities, purposes and goals. The table review elaborates Bomsdorf's emphasis on *"emancipated specification of task driven, role driven and content driven views"* [BB07]. This emancipation reflects the task modelling aspects of the metadata encoding framework and draws upon the modelling processes as defined by Paterno [P99]. In addition to that, this aspect of the modelling highlights task models' affiliation to *"usage oriented processes"* leaving system processes and domain orientation to the business process model [BB07].

Therefore, emancipated representations of interactions, roles and goals require relevant guidelines for processing whilst acknowledging the focus of concept tasks towards user goals and process models on task assignment. In light of which a task model is associated with the decomposition structure of the framework in the form of hierarchical tree notation as opposed to a sequential programming oriented process model.

	Process Oriented	Content Oriented
target groups	known users	unknown heterogeneous users
purpose / goals	execute task	search, browse and explore information
primary subject of design	functionality and access via a user interface	information and access via web pages
documentation	handbook	intuition
central paradigm	interaction	navigation
state information	meaningful state: task/interaction completion	stateless: current position
control	system control	user control
interactivity	complex	simple
metaphor	direct manipulation	navigation
genres	isolated dedicated applications	interlinked applications
basic design principles	usability	user experience

Table 3.5: Interactive characteristics comparison according to Bomsdorf [BB07]

As such collection of three methodical approaches summarized by Paterno [P99] appear relevant for modelling digital archiving tasks namely:

- ***Hierarchical Task Analysis (HTA)***
  Logically structured descriptions of the rudimentary set of activities in multiple levels and the numerical order of performance [P99]

- ***Goals, Operators, Methods, Selection rules (GOMS)***

  A pioneer model for systematic user interface design based on the description of the usability goals to be achieved [JN94] and hence the cognitive Human Processor Model with the perceptive, motor and cognitive interacting subsystems [P99]. The model is characterized by a set of memories and processors listing probable user goals, sub-goals with operators in addition to as set of selection rules as principles underlying their behaviour.

  - Goals
    resemble the targeted achievements to path to which is hierarchically described using operators
  - Operators
    defined as “elementary perceptual motor and cognitive acts” [P99]
  - Methods
    composed by users with the support of *“sequences of subgoals and operators used to structure the description of how to reach a given goal”* [P99]
  - Selection Rules
    serve as decision aids and indicators of appropriate methods and alternatives necessary to accomplish target goals.

  Nielsen [JN94] outlines the weakness of GOMS as a theoretical approach with limited *“error-free performance by expert users”* however acknowledging their suitability for the analysis of user interfaces for human-computer interaction.

- ***User Action Notation (UAN)***

  Generally most task analysis and modeling approaches are hierarchical in nature breaking down higher level tasks and goals into subtasks and subgoals subject to further decomposition. This logical hierarchical structure is the binding factor between task action analysis approaches GOMS and HTA described above with the UAN approach illustrated by the exemplary Fig 3.4. Despite the fact that no user interfaces are compelled to formal specification [JN94] UAN was developed specifically to communicate design. It resembles a natural language notation

combing classical software engineering process oriented notations and user action descriptions to represent user interfaces as quasi-hierarchical structured asynchronous tasks. The tasks are sequenced independent of each other with temporal relationships between them being described by a set of operators and structured as follows:

- task disintegration with respective descriptions of the temporal relationships among asynchronous tasks [P99]
- specifies task association description in the table columns
    - user action
    - system response
    - interface state modifications

Task: Withdrawing Cash

User Action	Interface Feedback	Interface State
~[Withdraw] V^ ~[Amount] V^	Display (Possible amounts) Provide (AmountCash)	CurrentService=Withdraw Account=Account-Amount

Fig. 3.4 User Action Notation

The task model of the metadata creation framework is based on judgmental task data [EM79] aligned to the formative evaluation and its associated results. The focus is indeed on the record collection tasks representing the main system user activities elicited from the aforementioned evaluation. The conceptual separation of record collection and structured tagging places the encoding dominates the high level task allocation. The primary tasks are summarized in the high level task frame and the associated decomposition illustrated in the figures 3.5 and 3.6 below:

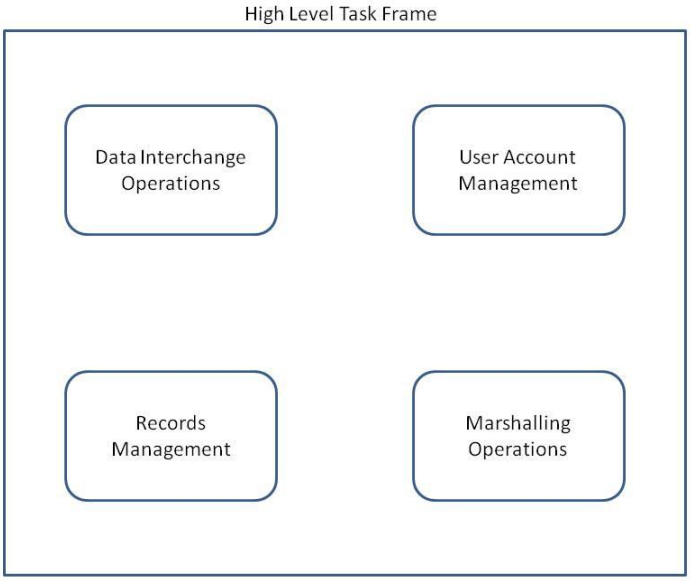


Fig. 3.5 High Level Task Frame

A decomposition of the aforementioned task frame reviews the system user oriented tasks in preparation of their consideration in the user interface design. The user interface design draws upon the same design and usability requirements defined by judgmental task data and system user observation and illustrated by the formative evaluation.

The entire task analysis and formative evaluation constitute a system design feedback correcting and improving design, usability and implementation issues elicited during the design phase.

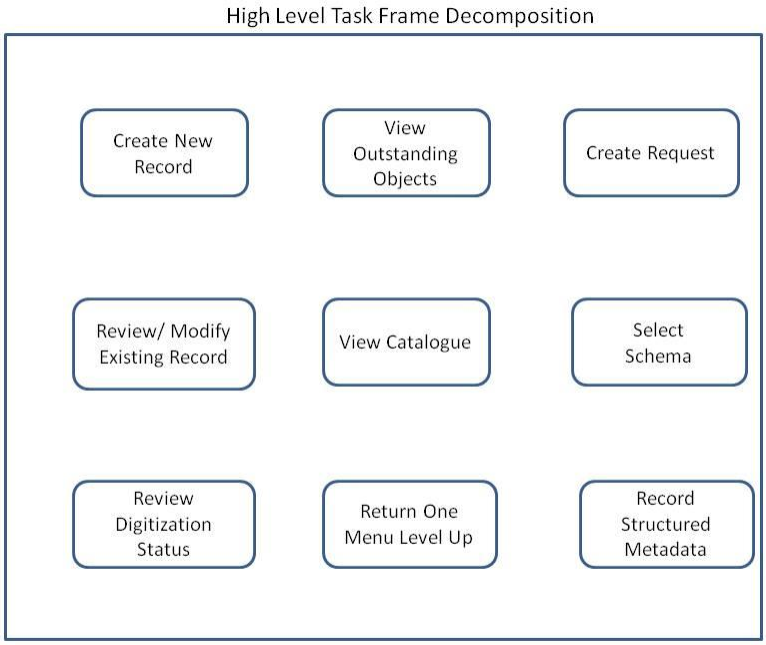


Fig. 3.6 High Level Task Frame Decomposition

3.2 Data Model Requirements

The analysis of the metadata creation framework described in the preceding chapter reviewed the management of structured heterogeneous XML metadata based on the semantics and goals of digital preservation as a primary function of any metadata creating system. According to the analysis the conceptual role of the framework is to mediate between business objects and the semantic elements for XML data preservation. However these mediations represent data relationships needing formalization, constraints and semantic representation whilst encouraging a shared understanding of the business data by the participating description schema. This formalization and its associated documentation constitute the data model of the framework and are tasked with the definition and the structural organization of the data format. This subsection looks at and analyses the requirements of the data model for the metadata creation framework looking at the definition of metadata constraints, the resultant entities and the relationships between them. The model resembles the data level requirements of the metadata creation system illustrating the boundary role of representing metadata structures in a way understood by the designated user communities [DE10] as required by the preservation needs discussed in the previous subsection. The data modelling approach is categorized into the following subsections:

- ***Conceptual Data Modelling***
  Identifies high-level semantics represented by associations and relations between different entity classes focusing on the overall system concept.
- ***Enterprise Data Modelling***
  Addresses lower level business function specifics
- ***Logical Data Modelling***
  Depicts a semantic specification of the business functions implemented either on class objects, XML elements including their attributes and relations.
- ***Physical Data Modelling***
  Illustrates database related aspects of the data model including database table structures and storage implementations.

3.2.1 The Entity Model

Entities for bibliographic records can be defined in line with the Functional Requirements for Bibliographic [FR08] records. These requirements have been translated to suit archives and scholarly collections in the RNA (Regeln für Nachlässe und Archive) [WK10]. These rules and recommendation are the guidelines for the definition of entities to be modelled and used within the framework of this dissertation and form the basis for the data and constraint models governing the metadata creation and interoperation for digital archives. The entities represent the classes of preservation metadata hand-generated by digital archivists and realized using XML meta-language elements and attributes, in other words tag classes. The entity class categories can be described as

- ***Product entities***
  The entities in this group reflect the spectrum of intellectual products of interest to users of a digital archiving system. Whereas the entities ***work*** and ***expression*** reflect upon content, ***manifestation*** and ***item*** reflect upon physical objects.
- ***Responsibility entities***
  The second group of entities handles aspects of content responsibility, authorship and custodianship represented by the elements ***person*** and ***corporate body***.

The entities in this group relate to those in the first group as illustrated in the figure 3.7 below:

- ***Subject entities***

  This group of entities constitutes *subject* entities as attributes of the product entities. The subjects include ***concept***, ***object***, ***event*** and ***place***.

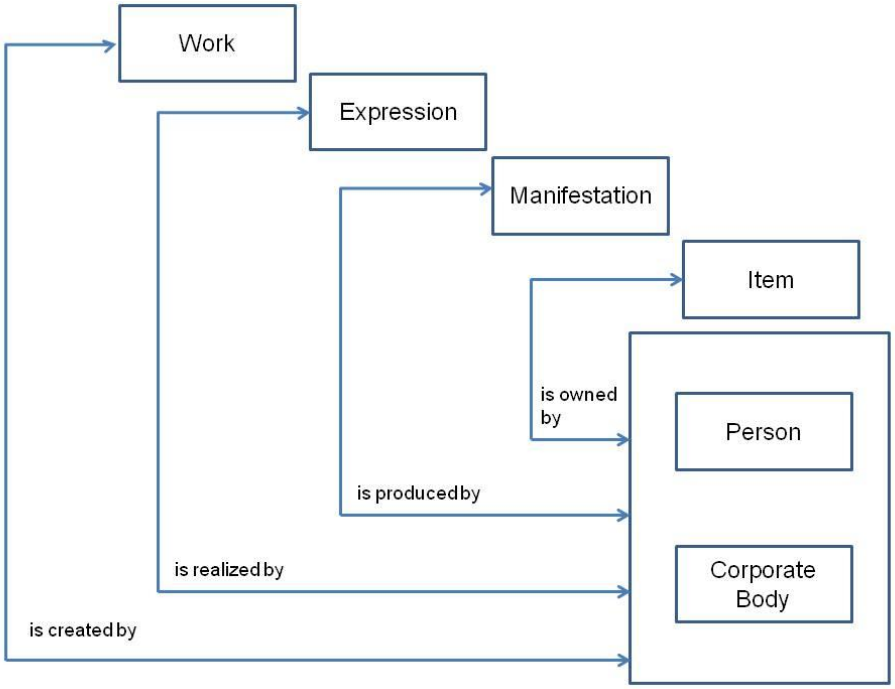


Fig. 3.7 Bibliographic Functional Requirement Entities [FR08]

3.2.2 Open Archival Information System (OAIS)

In chapter 2.1.2 bibliographic metadata and descriptive structures for digital archiving as internet based resources were introduced together with an illustration of the metadata harvesting model based on the Open Archival Information System reference model [BB02] [NH09]. The OAIS reference model is the subject of description in this subsection with its recommendations towards the definition of the concepts, terminology and elements associated with long term preservation and digital archiving. The reference model determines the functional entities of a digital archive whilst serving as an abstraction of the key concepts of long term preservation in digital archiving illustrated by a simplified framework. This framework serves the understanding of the necessary archival concepts key to long term preservation.

In addition to providing the reference for comparison between existing archival data models, architectures and operations with proposed standardized strategies and techniques, the framework resembles a gateway for the participation of *"non-archival organisations"* [BB02] in the standardized preservation process.

The OAIS model neither specifies a design or a data model nor an implementation of any kind; instead it identifies functions and responsibilities consistent with long term preservation within an OAIS environment. This OAIS environment together with the associated notions defining information its packaging and variations of the latter coupled with the defined functions and responsibilities resemble the open archival information system. The archival system as an ensemble of the OAIS concepts of environment, information and high level interactions outline the OAIS reference model. As such the reference model issues the guidelines for long term preservation outline by the OAIS notions bundled within a framework whose aims and purposes are summarized below as follows [BB02] [NH09]:

- Framework for an increased awareness and an understanding of the notions of long term preservation and access to digital information within an archival context
- Conceptual provisions for the effective integration of non-archival organizations into the long term preservation process
- Platform for comparison and standardization including terminological and conceptual standards in addition to architectural und operative standards
- Platform for outlining and for the comparison of long term presentation techniques and strategies
- Resemble a podium for discourse on digital information data models and their transition with time in addition to laying the foundations for the comparison of the respective models of the digital information data preserved by the archives
- Provide common grounds for the enhancement of further notions of long term preservation of information other than that in digital form
- Extends *"consensus on the elements and processes for long term digital information preservation and access "*[BB02] hence broaden the basis for interoperability and multiple vendor support
- Steers the identification and production of standards associated with OAIS

In the problem description of the dissertation, digital libraries and archives together with their associated metadata are described as serving as *"a novel mode of preserving, presenting and accessing cultural heritage "*. Should we compare this description with the summary of the OAIS illustrated above we quickly come to the conclusion that state-of-the-art digital archives and libraries are a form of information system seeking long term preservation of their information content. To this end such digital archiving systems should be conform to the recommendations of the OAIS and provide for interaction, interoperability and a common understanding on the basis of the reference model summarized above.

As such conformance is governed by section 1.4 of the reference model [BB02] requires the digital archives to support the OAIS concepts of an environment, information and high level interactions as outlined above. In addition to supporting the notion of OAIS, conformance further requires the fulfilment of the set responsibilities mandatory to OAIS however with the possibility to discharge some of these responsibilities. An illustration of an OAIS environment is illustrated by Fig. 3.8 below followed by a summary of the responsibilities necessary for OAIS conformance.

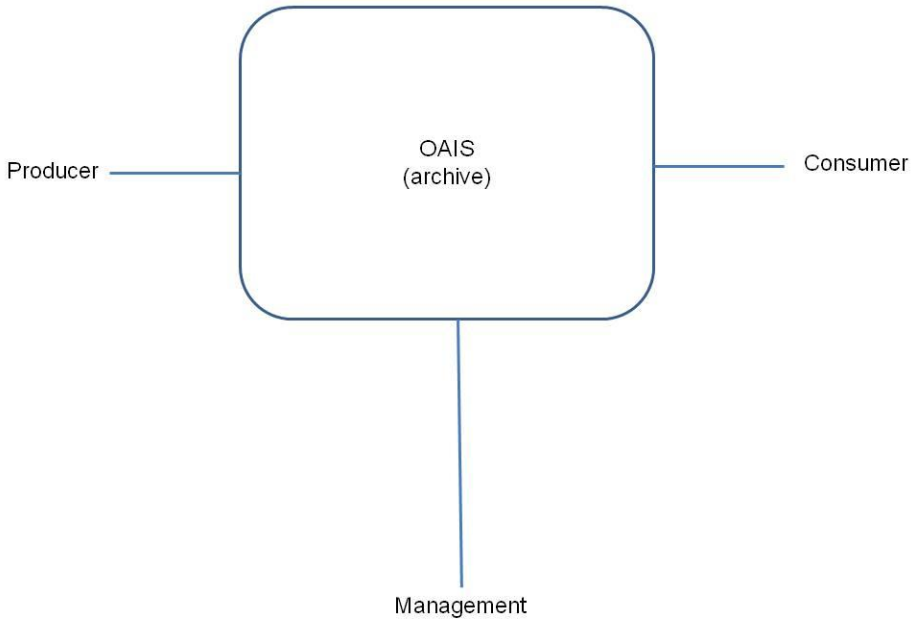


Fig. 3.8 OAIS Environment [BB02]

OAIS Concepts

As already mentioned in the previous subsection the prerequisite for OAIS conformance in addition to the responsibilities lies in the support of the key OAIS high level concepts environment, information and external interactions. These concepts form the basis of long term preservation functionalities and entities determined by the reference model and culminate in the functional and information models as well as the information package transformation necessary for conformance. The proliferation of computer processing, digitization and digital media has seen the necessity of information preservation activities previously common only to traditional archives. However, because digital information is easily lost or distorted [BB02] coupled with the rapid pace of developments in computer technology effective preservation is necessary. Effective preservation as sought by the OAIS requires principles as those outlined by the OAIS concepts, the concepts can be summarized as follows:

- OAIS Environment

  Consisting of the actors providing information for preservation – producers, managing and administrating the OAIS archive – management and finally the consumers interacting with the archive to find *"and acquire preserved information of interest"* [BB02]

- OAIS Information

  The central concept of information is key to OAIS and defines *"any type of knowledge that can be exchanged, and this information is always expressed (i.e. represented) by some kind of data"* [BB02] The system should be accompanied by a **Knowledge Base** allowing the reception of information in addition to **Representation Information** elaborating dictionary and grammar information.

  All in all the OAIS notion of information conforms to the semiotic pragmatic information in chapter 1.2 on structured encoding relating structure, content and impact. In other words environment participants interpret the data archived in the knowledge base using the representation information to yield information as shown in Fig. 3.9 below.

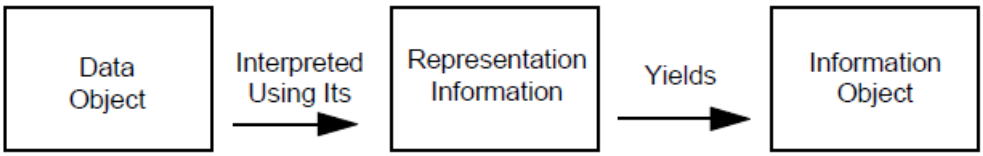


Fig. 3.9 OAIS Information Object [BB02]

In summary the desired information object subject to preservation is dependent on the identification of the **data object** and its associated representation information to achieve an effective long term and structured preservation. This notion of data objects plays a central part in the abstract creation of XML metadata via a graphical user interface. As already mentioned in chapter 1.2 XML documents are defined as *"the description of a class of data objects"* [KST02]. Furthermore, the separation of data processing from data storage complemented by this principle of data objects forms the basis for mapping the abstract metadata to XML as illustrated in chapter 4.

With the environment and the information in place, the reference model recommends a discrete transmission of information from the producer to the archive and from the archive to the consumer and defines the **Information package** illustrated in Fig. 3.10 for this purpose.

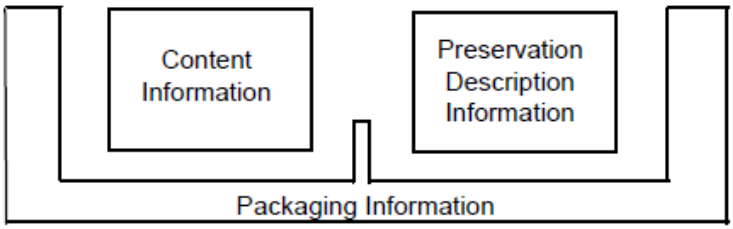


Fig. 3.10 OAIS Information Package [BB02]

- OAIS High level external interactions

  The participants of any OAIS environment operate as interacting functional entities of the OAIS archive with the high level external interactions modeled as management interaction data flows; producer and consumer interaction data flows. The figure below illustrates the view of the high level external interactions [BB02]

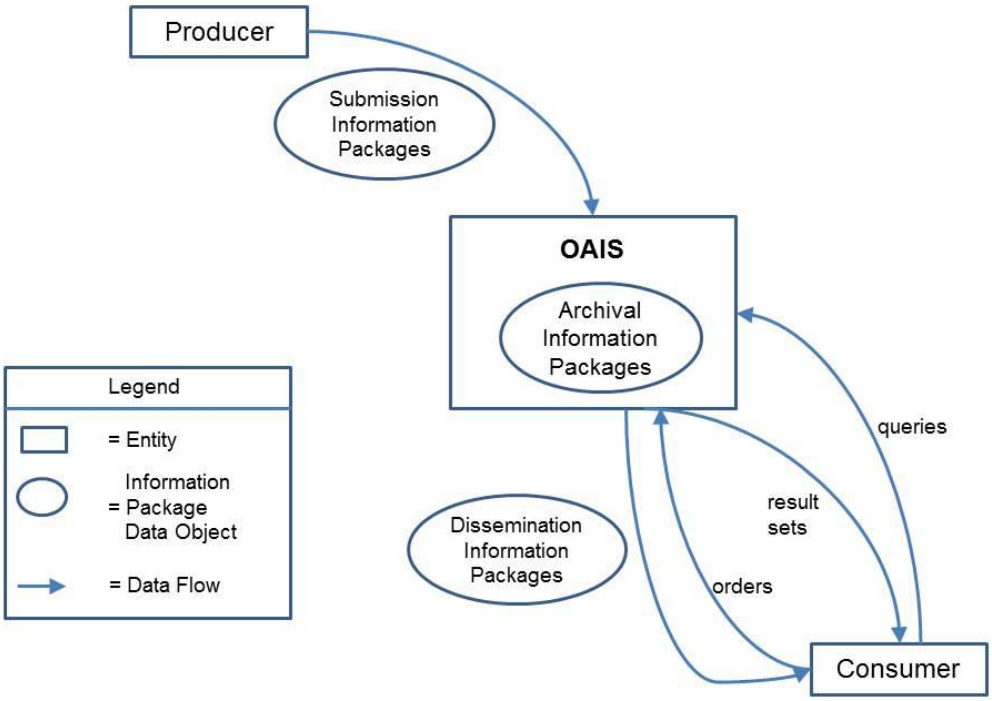


Fig. 3.11 OAIS Responsibilities [BB02]

In addition to the concepts, an OAIS archive is defined within the framework of the concepts illustrated in Fig. 3.11 above as *"one that intends to preserve information for access and use by a Designated Community"* [BB02] whilst fulfilling the OAIS Responsibility requirements. For the digital archiving community the information to be preserved and the functional entities are defined by the entity model based on the Functional Requirements for Bibliographic Records [FR08] and RNA [WK10] described previously. Comparing with the OAIS concepts above these functional requirements and hence the entity model show similarities and hence conform to the OAIS reference model. The entity relationship model of the entity model illustrated below outlines the relationship between the information objects and culminates in the subsequent data model for the metadata creation framework. It is therefore imperative to specify the responsibilities recommended by the OAIS model and match the RNA and the bibliographic functional requirements to these. In addition to that the data model of the metadata creation framework and hence the metadata as contents of a knowledge base must also be in line with the information package model contained in the OAIS recommendation. As a consequence of the above analysis we now summarize the responsibilities mandatory and discharged necessary for conformance with the OAIS and hence necessary for the tasks

associated with the metadata creation framework which is subject to this dissertation work.

- Negotiate and accept information from producers
- Obtain control of the information sufficient for long term preservation
- Determine the designated community and understand the information they provide
- Ensure preserved information is **Independently Understandable** without expert assistance
- Follow policies and procedures for provenance and authentication
- Accord the designated community access to the preserved information

3.2.3 The Data Model

A representative data model of the metadata creation framework considers the digital intellectual entity object as the nucleus of all description activities. The figure 3.1 below illustrates the object entity-centred data model in orientation with the ISO 13407 for data processing. In this model the object entity role is central to the coordination of descriptions of the digitized texts, images and records of concern. The object context refers to the reflective descriptions in the circumference of the digital objects as administrative and header metadata containing references to authorship, provenance, ownership and general context oriented texts. Descriptions of text and or related imaging complement the facsimile also defined as images traversing the heterogeneous model back to the original object document entity and its record. The model generally assumes a textual orientation in the description of the context. On the other hand the information data model illustrated by the taxonomy in the figure below shows orientation towards the OAIS information model and the associated information package transformations which resemble the lifecycle of the digital objects preserved in the digital archive. The information builds upon the OAIS concepts to describe the types of information managed and exchanged by the archive [BB02]. In other words the information model recommends conceptual data models for standardization in the information interchange process be it inter-archive or otherwise. In the case of the metadata creation framework, the information is question is metadata by nature and is the product of the interpretation of descriptive data using standardized schema as representation information. The taxonomy illustrates the hierarchical data structuring and the heterogeneity underlying each metadata object.

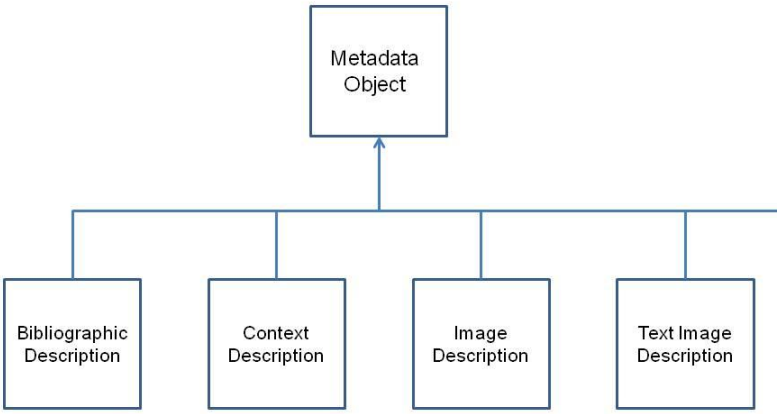


Fig. 3.12 Framework Information Model

In addition to the overall entity relationships, the refined data model of the metadata creation framework reflects the structure of the descriptive entities created as metadata and their storage in a persistency consist with their relational nature. To illustrate this phenomena the data model is refined to include the table structures in which the metadata will be stored and the relations between the tables in addition to possible associations as illustrated by Fig. 3.12. In general the model defines and describes the columns of the proposed tables of the respective relational database resembling in our case the structure and category of metadata to be created and hence the relevant java class structure. These columns are then filled with row values identified with the help of a primary key and constituting the descriptive elements collected as bibliographic metadata and to be encoded as the content of XML entities in the respective structured XML archival documents. The entity-relationship model defined as the bibliographic functional requirements dictates the structure and contents of the relational tables for the metadata creation framework. The resultant table for Person entities can then be summarized as follows**:**

Autoren-nummer	Autorenname	Beruf	Lebensdaten	Wirkungsort	PNDNummer
001	Altmann, Salomon P.	economist	*1878 + 1933	Mannheim , Freiburg i. Br.	11629518X
002	Barth, Paul	philosopher	*01.08.1858 +30.09.1922	Leipzig	118821326
003	Cassirer, Ernst	philosopher	*28.07.1874 +13.04 1945	Berlin, Hamburg	118519522

Table 3.6: Person metadata as contents of relation database table

gnd:foreName	gnd:surname	rdaGr2:professionOrOccupation	rdaGr2:identifierForThePerson
Salomon P.	Altmann	economist	11629518X
002	Barth, Paul	philosopher	118821326
003	Cassirer, Ernst	philosopher	118519522

Table 3.7: Person database table containing metadata categorized according DNB-RDF tags

The table 3.7 above resembles the set of description tags illustrated in the RDF section in the figure above and can therefore have its columns redefined in accordance with the RDF tags.

Whereas the gnd:RDF given metadata categories structure the refined data model, the extent to which they go will not be the subject of the dissertation, hence the proposed data model refined or otherwise of the metadata creation framework focuses on the core information required for digital archiving.

Returning to the relational nature of the metadata in for digital archive entities and their data model tables finds us in a position needing an elaboration of the associations between metadata entry tables within the data model of the framework.

To this end we note the entity relationships between the groups *Person/Corporate Body* and *Work/Item* as intellectual products seeking preservation. The resultant table associations are illustrated in UML above reflecting the structure and hence representing a refined data model of the contents of a digital archive collecting and encoded using the XML creation framework described in this dissertation.

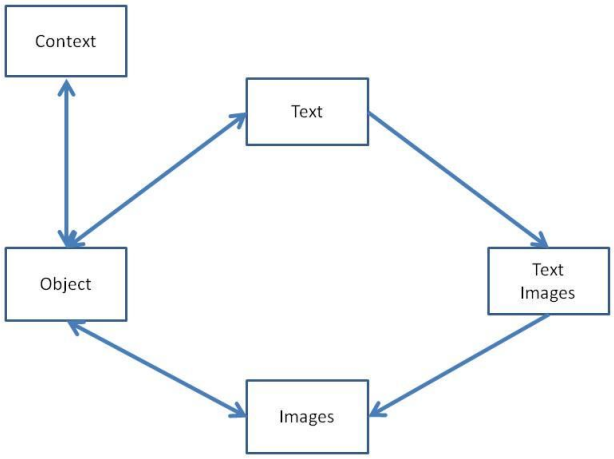


Figure 3.13: Metadata Creation Framework Description Data Model

The data model of an XML based metadata framework as illustrated in Fig. 3.13 above automatically emphasizes the XML concepts of separating the user interface from the data. However within the context of the archival metadata serving as a preservation mechanism, this notion is complemented by the principle OAIS concept of information being a combination of data and representation information. In other words, the data separated from the user interface by XML is only meaningful and preserved when representation and hence structural information is added to it. In other words the metadata created using the framework fulfill their function as descriptive, administrative or structural information when they are in possession of an association with some representation structure in this case an XML Schema. The RDF DNB XML record above illustrates this phenomenon giving an insight into the Resource Description Framework representation of the bibliographic metadata on Salomon Altmann. This exemplary description shows the separated (meta) data derived from the tables of the relational persistency, the structural representation information of the RDF Schema informs us of the role and the meaning of the data stored within the tags. Provence and preservation are then a matter of the description schema vocabulary with the indexed person being further identified via the technical P(G)ND number:

<gnd:variantNameForThePerson>Altmann, S.</gnd:variantNameForThePerson>

<rdaGr2:indentifierForThePerson>(DE-5882) 11629518X</rdaGr2:indentifierForThePerson>

Matching the semantic resource descriptions to the OAIS reference models reveals the OAIS concept of an **information object** [BB02] comparable with the notion of business objects in java. This concept is best visualized in the UML diagram shown in Fig. 3.14 below relating objects and information.

A closer look at the UML illustration reveals relational similarities between this model and the entities it represents with the entity relationship illustration of the Bibliographic Functional Requirements Entities discussed in the preceding subsection.

In addition to that, further similarities to the description data model of the metadata creation framework can be seen. These similarities may be attributed to the semantic representation of the objects in question and their descriptions. The semantic representation resembles a structured description of the content, structure and presentation of the archival objects regardless of their state. As a result the representation resembles a description overlap in the metadata implemented as representation information and necessary for interpretation as structural, content and context and hence preservation information. The entire information structure can be illustrated with the help of an n.-tier model view architecture encompassed within a framework interface with each n-tier assuming a data or representation information collection role suited to the OAIS notion of an information object. The physical object then resembles the physical archival artefact, the digital object assumes the role of the digitized object capable of further expansion into different digital modes but all viewed as the model tier. The data object together with the semantic descriptions as the representation information deal with the structure and hence the data control tier also responsible for interpretation. Finally the information object as the product of the process of integrating the data object and the representation information (the structural standardization) constitutes the view tier. The OAIS information object UML [BB02] is illustrated below with its identified relationships compositions and aggregations.

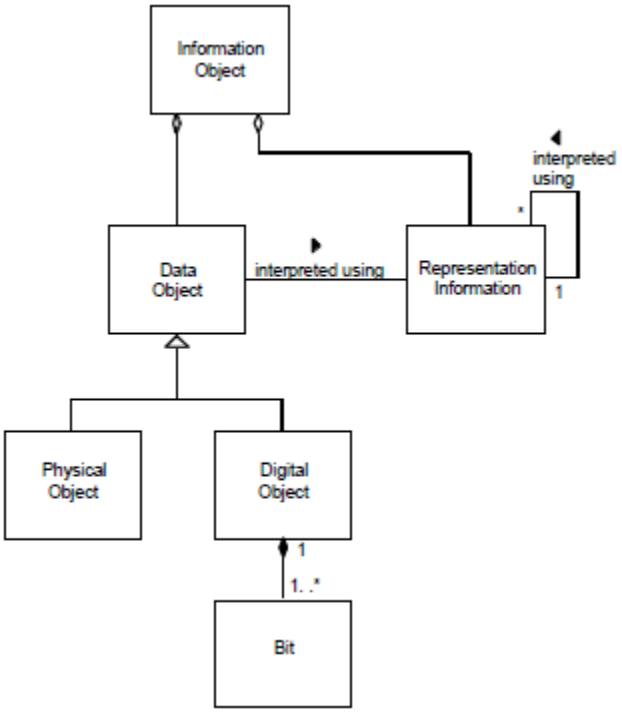


Fig. 3.14 OAIS information Object UML [BB02]

At this stage of the dissertation it is clear that representation information is the product of composite digitization of archival contents to be preserved in digital form. It is also clear that the digitization and retrieval processes are focused on the metadata as representation information which can be interpreted independent of technological developments and barriers on the basis of the description of the description information contained within a preserved information object. Consequently representation information as illustrated in the figure above is seen as being multi-tier in nature with sometimes complex inter-tier relationships between higher and lower level meanings. The illustration of the digital object shown above runs down to bit level supplying the bit sequences of the lowest level with **structural information** *"describing the format, or data structure concepts"* [BB02]. Such an OAIS representation of structural information yields common data types and aggregates common to mainstream computing such as numbers, characters, arrays, tables and pixels. Representation information necessary for human interpretation of digital objects and automated processing in the bibliographic sense is often accompanied by a language for expressing the structural information which it in fact complements. This semiotic description of the data types of the structural information is referred to as **semantic information**.

3.3 User Interface Requirements

One of the goals of this dissertation predefined in chapter 1 is the reallocation of selected encoding functions from human beings to software whilst maintaining data collection by human users with the support of a user interface. This subsection will analyse the requirements of such an interface and the associated functions prior to its implementation in the metadata creation framework. The analysis respects the definition of an interface as a *"the set of all signatures defined by an object's operations and describing the set of requests to which an object can respond"* [EG95] and is accompanied by an analysis of user tasks in the data collection process of the system. The associated formative evaluation, to be discussed in chapter 6 closes the interface system loop as feedback flows into the development and design of the user interface. Generally, the criteria for a good interface include [AG05]:

- *Intuitive interface*
- *Flexibility across platforms*
- *Confirmation of function completion*
- *Underlying protocol benefits*

In the case of the metadata creation framework, the primary purpose of the interface is to enhance usability in the metadata creation process. As such the user interface requirements are dictated by the user tasks as determined by the task analysis of formative evaluation discussed in the preceding subsections of chapter 3 and further dealt with in chapter 6 on evaluation. In other words the user role and the role defined user input constitute the primary criteria for the design of the interaction interface and hence it's requirements. In the case of OAIS conformance these roles and hence the user interface requirements, can be elicited from the principle high level data flows of OAIS archival operations excluding administrative activities such as accounting and billing. These data flows illustrated in the figure 3.15 below resemble the metadata creation process described in this dissertation following a lifecycle from the producer to the archive and from the archive to the consumer [BB02].

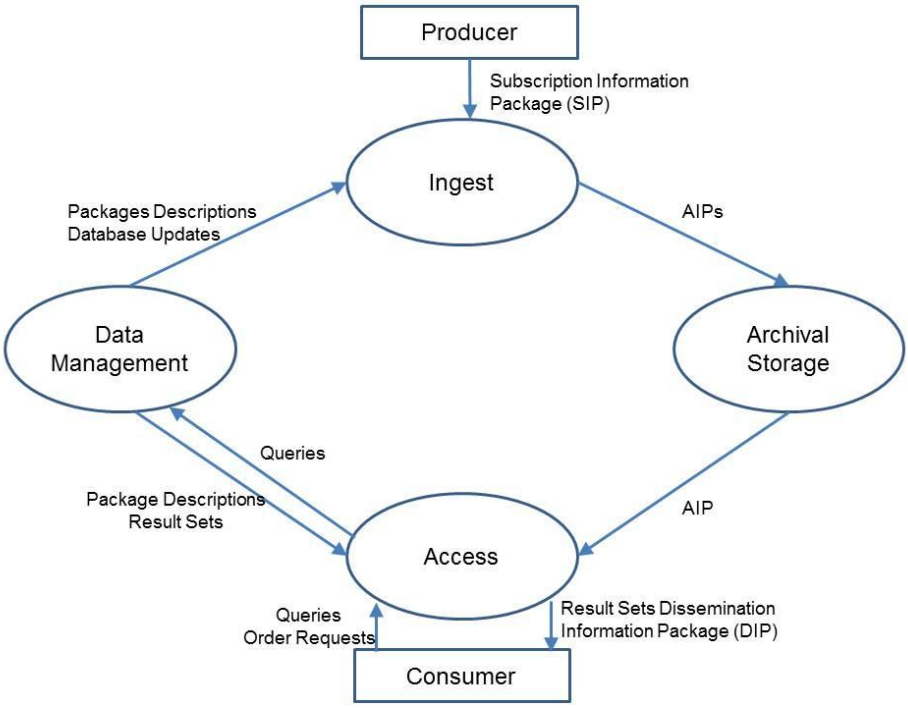


Fig. 3.15 OAIS Archive Lifecycle [BB02]

The central elements ingest, data management, archival storage and access illustrated above resemble the key interfaces enabling user interaction with the archival system within the framework of the defined associations. These key interfaces illustrate the elementary structure of the communication with the metadata framework which in turn hosts the structured XML metadata creation process. As such the central elements and the elementary communication structure should, in their role as the foundations of the user interface reflect the principle aspects of the metadata creation framework namely its:

- Functions

  The system functionality must be illustrated by the interface as a transformation process with a starting and an end state complemented where possible by a “current state”. In other words the interface should determine the flow of the process and adequately communicate the process and the flow to the prospective user.

- Characteristics

  Given that the framework resembles a semiotic repertoire of units which in turn constitutes the defined system, these characteristics units need should be represented within the set of user interfaces for interaction with the system.

In other words the archival functions and characteristics described as elements of the OAIS lifecycle elaborate the user interface requirements of the metadata creation framework from a systematic and semiotic point of view. This view combines the notion of the user interface as the window for reflecting a systems characteristics and functions with that of information as a system function. McCormick et al. [EM79] describe the latter by respectively differentiating between its two assumed roles of a stimuli triggering a system process resulting in a physical output and the transmission or processing of information meant as input for a further system. In other words the user interface may serve to monitor and steer a *"control"* process or serve as an impulse for starting an information processing or transmission procedure. In the case of the XML creation process of the metadata framework, the latter definition applies and is in line with the OAIS lifecycle governing interoperability and preservation conformance. The individual user roles of the producer and consumer respectively illustrate the metadata creation process in comparison to their navigation as the product of a comprehensive preservation process by the consuming researcher. However, these system-oriented user interface considerations have to be consolidated by adding on the system mediation aspects and criteria key to the dissertation question.

3.3.1 Mediation Process Interface

Now the classification of the role of a user interface triggering information transmission or processing described in the previous subsection denotes the metadata creation framework and its interfaces as a mediation facility. This implies that from an information point of view the system only mediates between the user and some further system and is hence controlled by the user who enjoys the power of either initiating or interrupting the metadata creation process and the digital preservation activities at large. In such cases *"the system control functions that human beings perform require the exercise of a wide range of human mediation functions"* [EM79].

In other words the need for some form of usability is exposed and that this usability is dictated by the necessary human mediation functions classified McCormick et al. [EM79] as:

- (Human) Information storage
  Long term: "The learning that is required for performing the system functions"

  Short term: "Remembering for short periods of time information that is relevant to a specific operational situation"
- (Human) Information retrieval
  Recognition: a perceptual process involving the recognition of relevant aspects of the user interface

  Recall: the recall in particular in *"short time storage"* [EM79] of relevant

    - Factual information
    - Procedures
    - Processes
    - Sequences

Information processing

- Categorizing – key aspect of structured archives
- Encoding – extensible mark-up
- Interpolating – complement text with digital text images
- Transmission - METS
- Transformation – the XML metadata in framework are to be transposed across a set of heterogeneous standardized XML schema

Decision Making

- Selecting standards, choosing datasets, planning etc.

Control of physical response

- Exercising control over the desired output

All in all these mediation aspects resemble a usability scenario and therefore require a usability approach for determining the user interface requirements. Although usability and the usability heuristics associated with in are tackled in the chapters three and six, the basic necessities for a usable user interface for the metadata creation framework are illustrated here. Theoretical approaches to usability analysis for user interface design include the GOMS method described in chapter three. The basic GOMS method [JN93] involves a listing of the:

- Goals and subgoals:
  e.g. creating a new document record
- Operators:
  Cognitive or perceptual primitives accessible to users e.g. mouse events, minimized memorable names
- Methods
  Compositions of sequences of operations used by the users to achieve particular goals
- Selection rules
  Decision assistance in cases of multiple methods for the same achieving a common goal

The notion of usability is one of the key purposes related to the motivation behind the metadata creation framework as such this notion is further elaborated upon in detail in chapter six together with the usability heuristics for assessment and the user evaluations. Nevertheless, the user interface requirements of the metadata creation framework are determined by the framework lifecycle representing the participating functional units together with the mediation procedures required to trigger respective functionalities. Although the above mentioned characteristics immensely contribute towards the design and the implementation of the framework user interface, this interface serves a client server internet environment. However such an environment resembles an international audience and may be required subjection to interface standards. Nielsen et al. [JN93] describes an elaborate interface standardization summarized in the following subsection.

3.3.2 Standard User Interface

The background behind interface standards is in the broader sense the general need for consistency in user interfaces. Although this consistency is the subject of heuristic evaluation in general it is gaining significance in usability studies as a key heuristic necessary in today's modern world of rapid technological changes. With the learning effort constituting the key usability criteria *learnability* standardized interfaces contribute immensely towards these criteria enhancing *"the users' possibility to for transfer of skill from one system to another"* [JN93]. The resulting consequence is the reduction of training effort and cost. A further advantage is the minimal user support requirements complementing the set of user and hence in OAIS terms consumer benefits as listed below:

- Consumer Benefits
    - Low user training cost and effort
    - Improved productivity and reduced number of errors
    - Better user satisfaction
    - Less user frustration
    - Improved user satisfaction

In general user interface standards must specify plausible interfaces with the least usability problems not only to the users but also to the designers and developers designing and implementing the interfaces. In addition further positive aspect of user interface standards are in the visible developer benefits summarized in their general form as follows:

- Developer Benefits
    - Reduced maintenance costs
    - Design pattern reuse
    - Solid stable developments
    - Comparable standardization

For the metadata creation framework the implementation of XML Binding in the java domain represents an implementation of common graphical user interface standards. The java graphical user interface libraries Swing and Abstract Windowing Toolkit are common in numerous day-to-day electronic devices hence familiar to most users. In most cases users are not even aware which technology lye underneath their trusted interface or menu particularly those resembling the windows file system. The user interface of the framework is to be implemented as Swing classes with menus common to standard file systems common to any computer user including those classifying themselves as *"luddites"*. The user interface requirements are therefore based on Swing classes as the interface standard and further defined within the framework of the OAIS reference and hence the elements illustrated in the OAIS lifecycle illustrated above. The mediation processes and their requirements are dictated by the nature of the framework tasks namely metadata creation. As such the focus of the mediation requirements is on the producer element as the main user and the information processing aspect commencing with the record collection activity complemented by the persistency. Further mediation processes focus on the generation of the XML metadata from the data stored in the persistency as a product of the initial interaction with the framework. Furthermore, the generation of the XML documents hosting the metadata requires mediation processes for triggering the XML binding process.

As can be derived from the descriptions above, the decision to design java based interfaces is not a coincidence. The advantages of java as the foundation for the framework programming activities as part of the binding architecture are obvious. However, propagation of the framework as a client server system also benefits from Java's platform independency further support of the interface notions of "*flexibility across platforms"* and *"underlying protocol benefits"* mentioned previously.

In summary the user interface requirements go beyond the general criteria to cater for their role in the framework architecture which sees the implementation of the *"doctrine"* of separating content from the presentation format hence separating data from the user interface mentioned in chapter 2.1.5. The subject specific requirements do justice to the task at hand whilst being governed by the roles outlined by the recommendations of the open archive information system. These user interface requirements are concerned with associated roles of being a data *"producer"* or *"consumer"* and the respective tasks carried out by these roles.

In addition to the tasks specific interactions, the user interface poses as the interaction mediator providing for communication across the abstract encoding environment to be catered for by the user interface requirements. Finally the usability aspect also plays a major role in the design and requirements analysis of the user interface. Since the success of the entire framework and its use is dependent on its acceptance by the target users it is these target users who determine how a usable interface has to look like. From a software development point of view, this criteria can be attended to by analyzing possibilities related to standard user interfaces already common to and indirectly accepted by the target user groups. Therefore the user interface requirements include the need to integrate standard user interfaces to boost usability and reduce the development effort as already illustrated in the preceding subsections.

All in all the user interface requirements of the metadata creation framework can be said to be on the basis of the following criteria:

- Separating XML data from the user interaction as the presentation format
- Serving the OAIS lifecycle roles *"producer"* and *"consumer"* as the interaction medium
- Mediating the abstraction between the encoding and the producer as they interact
- Standard user interface considerations as usability guarantor

It is these criteria together with the standardized criteria introduced at the beginning of this chapter that constitute the overall user interface requirements of the metadata creation framework. Their implementation and relation to the rest of the framework architecture will be dealt with in the chapter 4 and 5 where the enterprise architecture is introduced and implemented. Furthermore, the underlying concepts behind each criterion build up the solution to the dissertation problem and are the key visible elements of the proposed approach.

4 System Design and Architecture

According to McCormick [EM79] system design and development *"refers to the various procedures and processes that are involved in designing and testing systems of all kinds"* however not in the form of one-size fit all but, instead specific to the objectives of the system to be developed. In other words an umbrella term *"that can cover a broad conglomeration of operations"* [EM79]. On the other hand Gamma et al. [EG95] refer to the *"general arrangement of objects and classes that solve the problem"* in addition to the dictation of the system architecture by the *"set of cooperating classes "* i.e. the framework. This chapter resembles an overview of the functionality of the proposed metadata creation framework being realized by arranging and relating the objects, classes and procedures identified in chapter 3. The functionality implements associated processes mediating between metadata classes and respective XML schema outlining the rationale behind the metadata creation process as a system. It covers therefore, a broad conglomerate of operations and procedures involved in designing the general arrangement of objects and classes cooperating to create digital archive metadata.

4.1 Concept and Methodology

In the preceding chapter, we analysed the requirements for creating heterogeneous metadata within the framework of an integrated digital archive and the associated tasks and objects. Subsequently, we now introduce an overview of the proposed framework and the system architecture enabling its realization by applying the described method of binding abstract metadata class representations to XML schemas. The system design outlines the rationale behind the components and interactions necessary for the fulfillment of the framework requirements pointing out the possible inline and external customization interfaces and mechanisms which can prove to be useful for extensibility purposes. The framework as a whole demonstrates how heterogeneous XML metadata can be abstractly created on a schema neutral platform whilst still providing for interoperability despite semantic heterogeneity. Whereas the metadata creation takes place on a neutral platform, the platform itself is to be developed in the best case within the reigns of a classical programming language. The advantages of an appropriate programming language lie in the amplitude of libraries and assisting packages reducing voluminous programme code

and simplifying the implementation of the framework. At the moment, the most appropriate programming language containing packages XML related processing as well as classical presentation and distribution architectures is the Java programming language. This language and relevant aspects of its architecture, application programming interfaces APIs and implementation make up the subject of the subsequent subsection.

Java Programming Language

The development of the framework proposed in the preceding subsections requires the selection of appropriate programming tools, in the best case offering extensive standard XML processing libraries and platform independency. The java programming language and its enterprise edition resembles such an appropriate programming language rich in libraries and concepts suited to multiple user and distributed access. Among the key java concepts are Enterprise Edition concepts consisting of the Client Server Model, Multi-Tier Architecture as illustrated in Fig 4.1 below and the Application Programming Interface API altogether representing the spectrum of web-based business application development. Whereas the client server aspects are of importance to the framework described here, it is more the Multi-Tier Architecture and the XML Application Programming Interfaces that are well suited to the abstract data collection and processing proposed for digital archiving in this dissertation. In addition to the java concept of data objects and the separation of data processing from data storage, the associated integration of XML and XML processing in Java's Enterprise Edition relays a sense of confidence towards the proposed multi-tier framework for creating heterogeneous XML metadata. The classical multi-tier model and its relation to the metadata creation framework are illustrated in the figure 4.1 and elaborated in the subsequent subsection. The figure 4.1 represents a summary of the multi-tier architecture notion outlining the specific roles of the individual tiers and the three core tier levels. Individual tiers may be implemented on different hardware and still interact with each other.

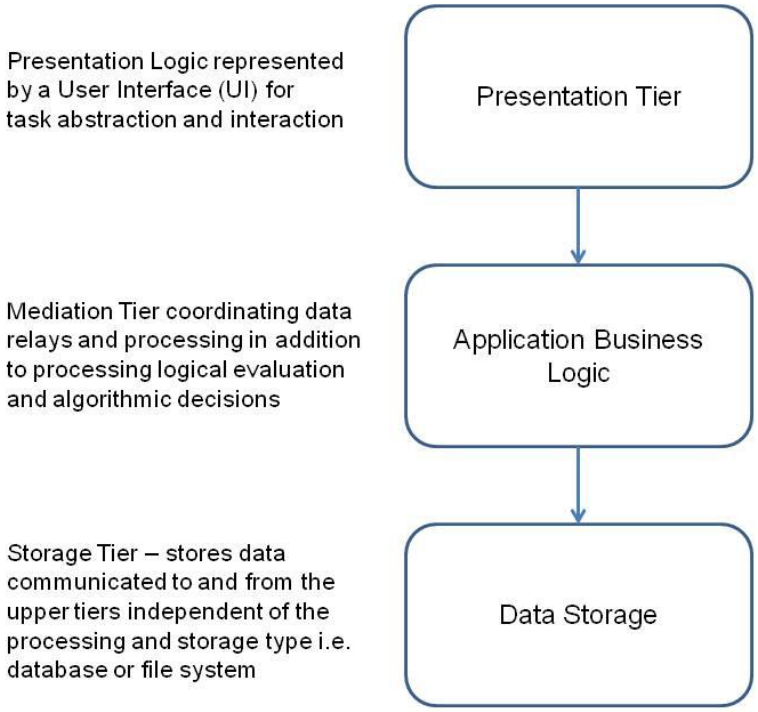


Fig. 4.1 Multi-Tier Architecture [RA05]

Multi-Tier Architecture

The general concepts of software architecture encompasses system organizational decisions relating system design, tasks and their interaction sequences within user interfaces and the layout of components and subsystems. These decisions are often closely related to the business tasks of the system being development and therefore tend to be or be oriented towards a software pattern. Andleigh et al. [AG92] realizes this referring to software architecture in general and database architecture in particular as layout of *"the components of the system, the services provided by each component, and the manner in which these components interact"*. This definition tallies with that of Buschmann et al. [BS07] acknowledging the *"set of significant design decisions"* including a *"description of subsystems and components of the software system and the relationships between them"*. The specification of these decisions and descriptions together with the way they collaborate illustrates a software pattern. The system architecture of the digital archiving business described in this work follows the path laid by the architecture of the predominant programming language JEE i.e. the n-tier architecture. This architecture supports the underlying principle of separating user the interface from

the data and its management as well as record collection from metadata tagging (encoding). The system organizational aspects further direct us towards the concept of components and the related java component model associated with the n-tier architecture and elaborated upon in the next subsection.

4.1.1 The Component Model

In general, the notion of components or the component model refers to the purpose oriented modularization of a set of functionalities and their encapsulation to be implemented as independent applications which can be assembled into an overall application development. In the case of the java component model(s), one differentiates between the server-side and client-side component model both representing isolated independent functionalities however serving different purposes. Both component models are of interest to our graphical user interface framework with respect to the GUI functionalities and the isolated business (record collection and encoding) functionalities. A general description of both models according to their purpose may be summarized as follows:

- ***Enterprise Java Beans Component Model***
  This server side component model together with the role model serve inter-process components associated with the n-tier architecture of the application as a whole. The most relevant components in the *javax.ejb* package of Java's EE architecture are the:
  - Application clients:
    implemented on the Client-Tier
  - Java Servlets and JSP:
    implemented on the Web-Tier
  - Enterprise Java Beans
    implemented on the Business-Tier

  The Business- and Web-Tier are implemented on the server-side as part of the JEE-API and hence subject to conformity tested during deployment. The figure below illustrates the n-tier architecture from a component point of view and forms the basis for the record collection architecture of our metadata creation framework.

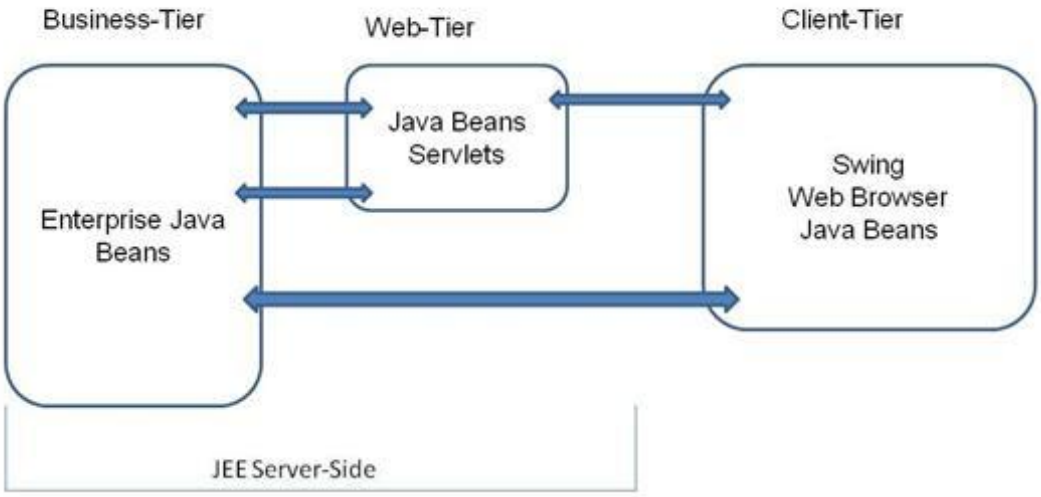


Fig. 4.2 EJB n-tier Component Model

- ***Java Beans Component Model***

  This model was designed within the framework of the original *java.beans* package as intraprocess components mainly for assembling graphical user interfaces (GUI) used for implementing and using client interfaces. However this model was not intended for distributed components and is seen to be more client-oriented. The figure below shows an illustration of such a distributed component architecture and its relation to n-tier architectures within a distributed client-server environment.

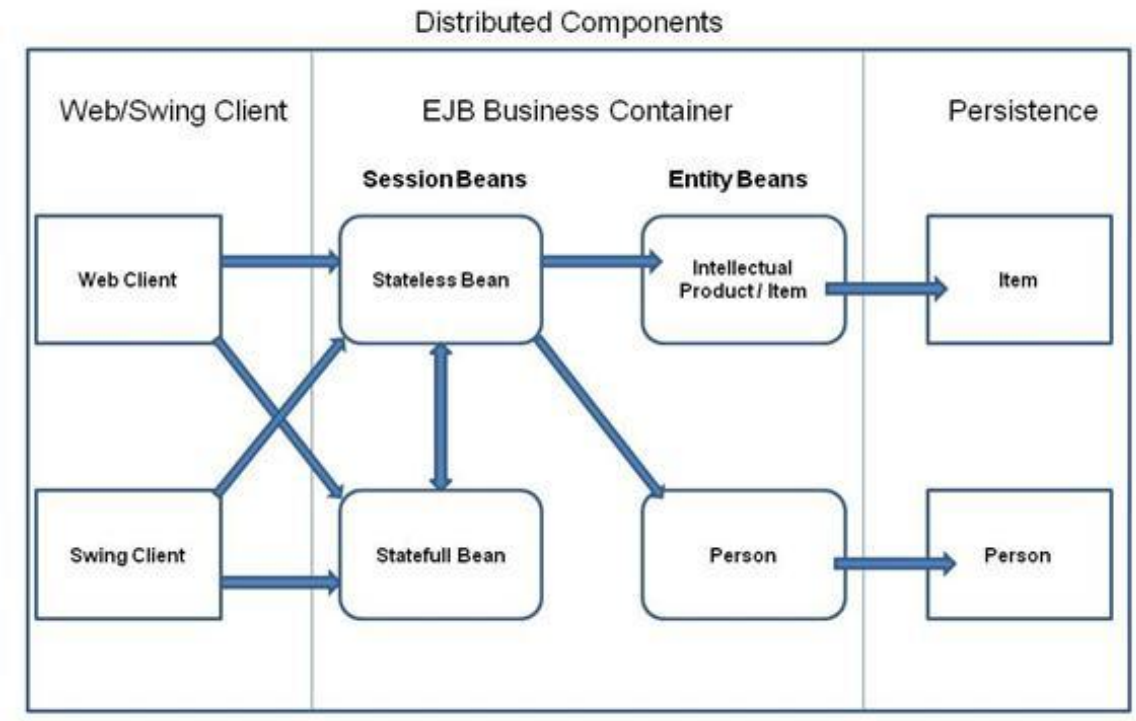


Fig. 4.3 n-tier Distributed Component Model

Java Persistence

The discussion on the n-tier architecture mentioned above has already introduced the concept of persistence in relation to the mapping of relational data into some storage media be it a database, file system or any other storage media independent of the media type. In Java the persistence tier consists of four subsections:

- persistence criteria
- query language
- object / relational mapping
- persistence API(as of Java EE 5/ Java 6)

This separation of application and storage serves the distributed processing of data objects within the framework of the independent client applications allowing multi-channel access and processing most suitable for interoperable systems. In other words persistency refers to data storage facilities resembling relational database architecture and accessed independently by the business entities of the middle tier of the respective java application [RA05] [BD07]. In this case data objects within the persistency domain

become *entities* represented by a table in the relational storage (database) unit whose rows correspond to an entity instance. Since the java applications are n-tier in nature persistency is managed via the middle tier either via the java beans as Bean Managed Persistency (BMP) or through the container as Container Managed Persistency (CMP) [RA05]. This introduces us to further characteristics of the middle tier namely, the java beans in either statefull or stateless session form or as entity beans as elaborated in the subsection below.

Entity Enterprise Java Beans

The business application logic of the middle tier in java applications is implemented via Enterprise Java Beans (EJB) incorporated within a middle tier container. Whilst the EJB beans encapsulate the business logic, the EJB container manages the "*system level services"* serving as an interface between the architectural tiers. The enterprise beans fulfil the actual purpose of the application and render their services and results to the client applications whilst simultaneously accessing the persistence [RA05] [BD07].

In our business case consisting of creating metadata as structured records of archived intellectual material, entity beans prove to be the logical choice for implementing the business logic. Given the nature of the record collection business, metadata reflect upon and represent objects within the business logic. These real objects making up the content of a digital archive are associated with persistence data stored and managed in the storage tier. With the distributive nature of the metadata creation framework, several clients can access any instance of an entity bean and use this to create, manage and maintain metadata records in the persistence. The classification of the enterprise entity beans is based on their responsibilities with respect to their persistence management structures allowing classification as either:

- Bean Managed Persistence
- Container Managed Persistence

4.1.2 XML Data Processing

One of the key concepts behind digitization is the associated machine readability and processing of the structured documents and their descriptions. Having this in mind, it is imperative that one looks at machine readability and processing of digitized structured XML data in general. Furthermore, it is also of importance to relate this processing to the current *"use case"* of XML encoded metadata to be collected using the framework designed here and their relation to the object oriented and schema neutral java digital metadata objects. Whilst XSLT and XPath remain the standard XML data processing languages, java object based processing is taking its place via the implementation of application programming interfaces commonly known as APIs for validating, transformation, generating and parsing XML documents. The common factor associated with the application programming interfaces has been the capability to read and interpret data from structured XML documents as an alternative to the Document Object Model (DOM), parsing XML data by creating representations via model interfaces or sequentially streaming the data.

Java XML APIs

Modern and interdisciplinary software development relies on standardised libraries and packages of the programming language and its development environment implementing the notions of code and pattern reuse in addition to enhancing efficiency in the development process. In our case where XML metadata are to created using java objects the use of such libraries to support Java⇔ XML is necessary and implemented with the help of Java Application Programming Interfaces API.

The standard basic Java APIs are Simple API for XML (SAX) and Java API for XML Processing (JAXP) however, of importance to this dissertation is the Java Architecture for XML Binding (JAXB) which will be elaborated on in the succeeding subsections. Generally API's are classified according to their data access levels as follows:

- Low Level API's
  Complex in nature, however grant access to the XML document's data and structure hence common for messaging and infrastructural tasks and are classified as follows [BM02]:

- Streamed data ⇔ SAX
  As one of the most common Java APIs for processing XML, SAX is an event based for top-down processing developed by an open source mailing list group XML-DEV. SAX sequentially parses an XML document and forwards the resultant events directly to the processing classes via the *Callback Method*. The callback methods include among others:
    - startDocument()
    - startElement ()
    - characters ()
    - endElement ()
    - endDocument ()
- Modelled data ⇔ DOM/JDOM
  This API follows XMLs hierarchical structure and models the tree structured set of XML nodes as a hierarchical structure in the persistence or file system.
- Abstracted data ⇔JAXP
  This Java API serves as an abstraction for the other two processing API models SAX and DOM/JDOM mentioned above. In web services JAXP is replaced by the remote procedural call RPC-API but more for interoperability purposes.

- High Level API's
  less complex in nature as java classes make use of "*the business purpose of the document rather than the data"* [BM02] with restrictions however, to less complex data processing and classified as follows:
  - Mapped data ⇔ XML data binding
  - Messaged data ⇔ Web services (Simple Objects Access Protocol SOAP)

Data Binding

Whereas the low level APIs summarized above read and traverse XML documents, their high level compatriots are data mapping oriented representing the XML documents as business-driven object classes [BM02]. This data mapping is generally referred to as data

binding and becomes XML Data Binding in cases where XML is the data source *(store)* [BM02]. Now, for the metadata creation framework the standard XML Schema represented by the schema documents (XSD) resemble the data source upon which the classes are defined justifying the reference to XML Data Binding. This reference to XML Data Binding together with the Java n-tier architecture and the graphical user interface defines the main concept and specifies the proposed metadata creation framework and its system architecture. As such we will now look at XML Data Binding architecture as well its role as an API within an n-tier Java application system and how each unit contributes towards the proposed metadata creation framework and user interface whilst respecting the data and entity models limited by the entity relationships specified by bibliographic and archival functional requirements described in chapter 3.

XML Data Binding

An interface and framework supported heterogeneous metadata encoding approach is based on the notion of a business driven record collection and description encoding infrastructure for digital archives. The main aim of the encoding process in the web oriented digital archiving environment is to generate structured XML documents qualified and understood within the framework of a standardized schema. This encoding process makes use of the concept behind the mapping of vendor neutral abstract data classes and variables of the record collection environment with the elements of the desired XML document structured with respect to the specified purpose oriented XML constraints i.e. XSD schema. This mapping generally referred to as data binding, represents a technique which allows the transformation of XML data elements into the object classes of common to object orientation. The main idea behind this approach provides for object relational record collection and maintenance whilst upholding the hierarchical structured documentation and preservation descriptions. The approach integrates data binding and record collection in a mediating environment for digital archiving offering a hybrid solution for heterogeneous metadata collection and associated structuring in XML documents. The abstract object relational record collection does away with multiple data entries of metadata relating to the same archival entities in favour of multiple constraints i.e. schema and element overlaps based on the same description entity. In exchange, the mediation between the entity object classes and the structured XML constraint elements and attributes, maps metadata descriptions with the respective schema structures resulting in

a heterogeneous scenario of multiple XML documents structured respectively on the basis of the same object oriented description data set.

In light of the hierarchical structure of XML documents and record collection tasks draws attention towards the object relational nature of the record collection and data entry tasks, highlighting the possibilities and advantages of entity centred digital archive metadata creation. XML Binding activities for digital archiving purposes in our research focus on the java programming language and its wealth of low and high level XML application programming interfaces (API). XML data binding in JAVA makes use of these low level APIs however concealing underlying details such as entity resolution and validation. The data binding process itself is encapsulated into a binding package consisting of:

- class generation
  involves the creation of instances of java objects on the basis of DTD or XSD schema constraints. XML java conversions can begin once the java classes are compiled, e.g. consider an XML element person with an attribute id and a child element surname resulting in the java class (Person) with the associated getId( ) method as well as getSurname( ) method.
- unmarshalling
  refers to the transitional process of converting XML documents into java classes commencing with a valid XML document conforming with the defined standard schema defined in the class generation section.
- marshalling
  resemble the key implementation of the proposed metadata creation approach, involving the conversion of java objects into an XML document representation. The validation takes place on the java classes before their implementation in XML hierarchical structures. The resultant marshalling process is laid out as follows:
  - java object and data validation
  - conversion of data objects into XML documents
  - storage of XML documents
- binding schema
  resembles a specification of XML constraint based class generation resulting in element based java objects i.e. <msDesc> has the java object MsDesc.

The specification handles respective name transformations and superclasses for generated objects modeled to the archiving business needs.

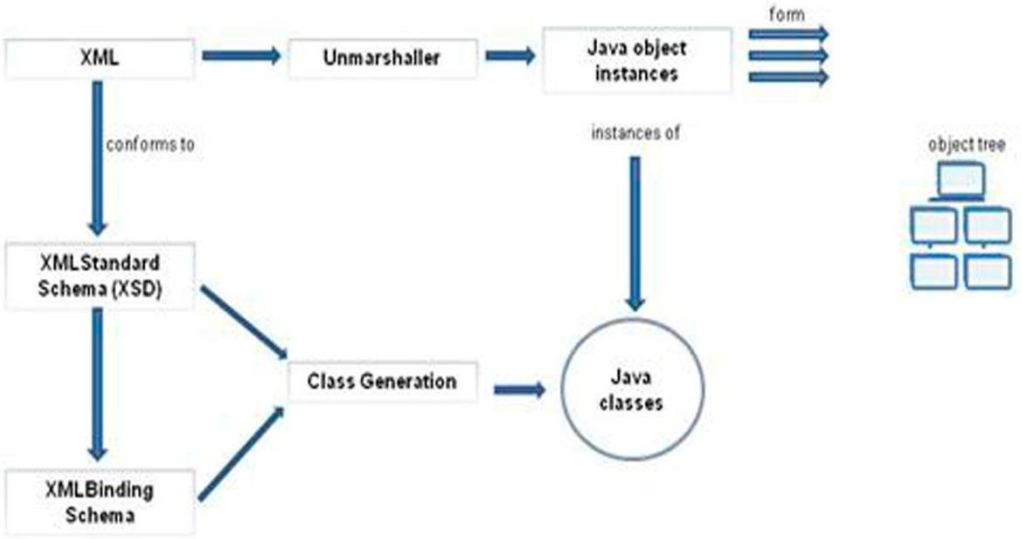


Fig. 4.4 JAXB Class Processing

XML Schema Patterns

A close look at the standard Java XML API's reveals their orientation towards reading data from and by parsing XML documents i.e. the low level processing of XML data hence Low Level API. The concept of XML Binding, a high level concept, on the other hand focuses on the incorporation of *"XML data and processing functions into Java applications"* thereby binding java objects and representations with XML elements and schema. The binding itself is guided by a binding architecture describing the interactions and interacting components and the generation of object classes and XML schema and elements. Object⇔XML element binding for XML and Java takes place within the framework of the JAXB.

4.1.3 JAXB – The Java Architecture for XML Binding

There are several data binding packages offering data mapping not only between XML and java but also SQL ⇔ LDAP e.g. Zeus and Castor, however all capable of XML binding in java enterprise environments. The notable integration factor between all these binding packages is the Java Architecture for XML Binding (JAXB) whose overview we are going to look at in this subsection. The general JAXB implementation consists of a:

- Schema compiler an XML described binding schema mapping the XML from the data store to the set of derived class elements as illustrated in the subsequent figures below.

- Schema generator concerned with the binding and hence mapping to a derived XML document on the basis of existing class elements with the mapping being described by the package annotations and in the case of JAXB, JAXB annotations.

- Binding runtime framework reflects the actual data binding during runtime providing for XML parsing reading (unmarshalling) documents for conversion to java objects and vice versa writing (marshalling) java into XML documents. These runtime operations include access and validation against the XML constraints specified by the XSD Schema or a Document Type Definition (DTD).

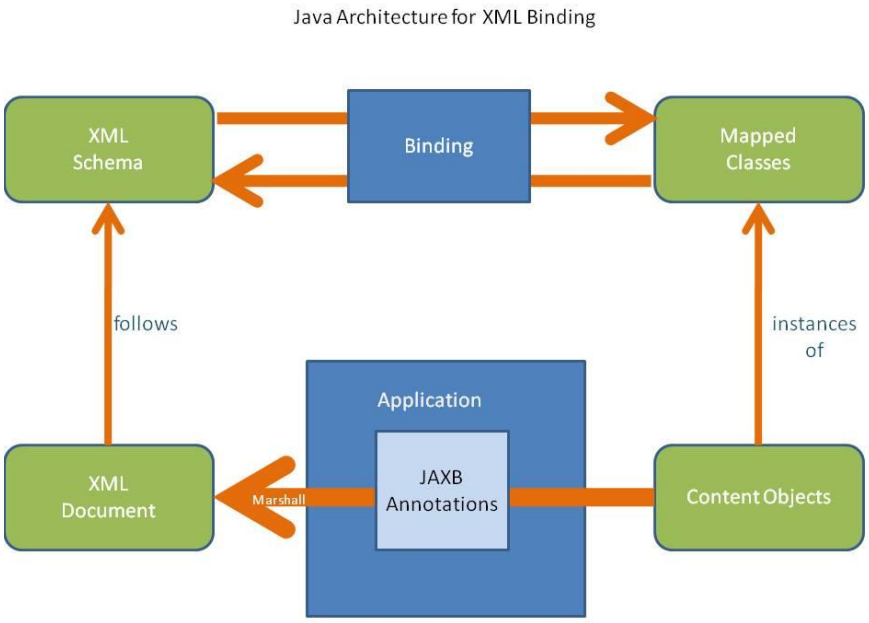


Fig. 4.5 JAXB Data Binding Process

Figure 4.6 [OM03] below shows an architectural overview of the JAXB application programming interface and summarizes the data binding processes described an illustrated in the preceding subsections. This architecture structures these processes allowing for the following data binding stages:

- Class generation
- Compile classes
- Unmarshalling
- Generate content tree
- Validation
- Content processing
- marshalling

The architecture illustrated in Fig 4.6 summarizes the functionality of Java XML-Binding using the JAXB infrastructure whilst complementing Fig. 4.5 above. The dotted arrows illustrate the unmarshalling process resulting in the structured objects and the annotated classes on the basis of which the XML Binding takes place. The Portable JAXB-annotated classes also serve the marshalling process enabling adjustments to newer XML schema

on the basis of the structured Java objects. The unit responsible for this process, the schema generator is also part of the architecture as seen in Fig. 4.6

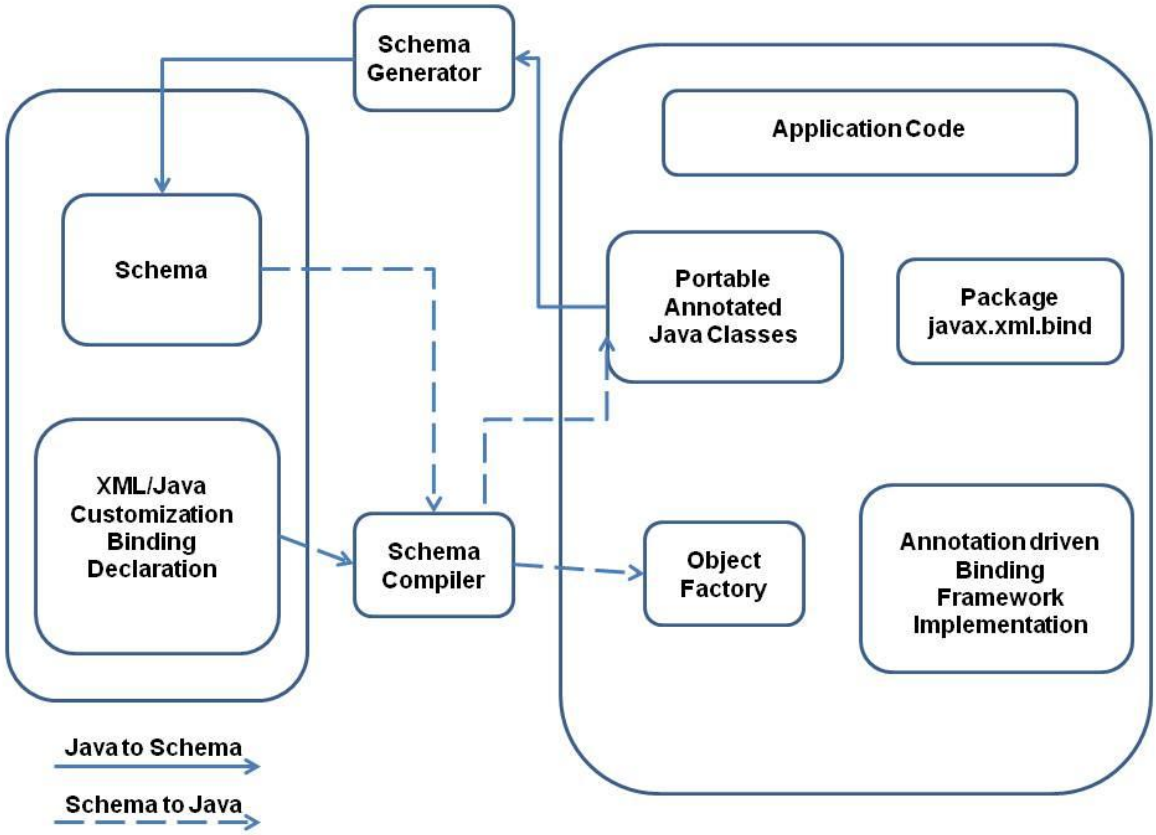


Fig. 4.6 JAXB Data Binding Architecture [OM03]

Summary

The JAXB architecture illustrated above enables abstract XML encoding without encoding knowledge and is doing access to XML data and it's processing without any previous or existing knowledge of the XML meta-language. It is this characteristic that makes JAXB suitable for the abstract metadata creation framework and its digital archiving encoding activities. Access to the XML document including XML Schema Document XSD is provided by the binding stage and results in the representation of the XML structure in Java format. The binding schema governs this schema representation in java by generating the set of java classes representing the XML Schema Document. Invoking the binding schema generates the set of classes and their associated interfaces with unmarshalling referring to the creation of the content object structure that maps the XML document format and structure. For the XML metadata creation JAXB provides for the

binding defined within the binding schema creating the content tree illustrating the target XML structure also achieved by using the Objectfactory before finally marshalling the content into an XML document. This process is the central notion to the swing interface implementation and the abstract structured XML metadata creation addressing the dissertation question. This possibility of actually creating a structured XML document using content accessed via java resembles a novice structured digitization approach. A combination of the JAXB architecture illustrated above and the model view controller architecture of the java enterprise environment results in a reasonable client server accessible framework for creating structured metadata. This framework implements the creation process without the necessity to learn XML and at the same implementing the XML *"doctrine"* of separating content and format respectively data and the user interface. As such the architecture and the framework approach are suitable for implementation in digital humanities.

4.2 System Architecture

The metadata framework for digital archiving resembles an interactive application requiring respective architecture suited to the task model and the associated interactions. The main task of creating metadata complements the system tasks of associating recorded metadata to XML structural standards. Whereas this transformation and XML mapping architecture has been described in the preceding section, the user interface architecture described here focuses on the dialogs and tasks related to each user's role as part of a semiotic framework. To this respect Paterno [P99] identifies different kinds of architectural models based on the tasks involved in the system to be developed summarized to include:

These basic models elaborate the principle architectures for logical input systems relevant to task, task pattern and interaction associated designs.

- Seeheim model – Model View Controller
- Arch Model
- Lisboa Model

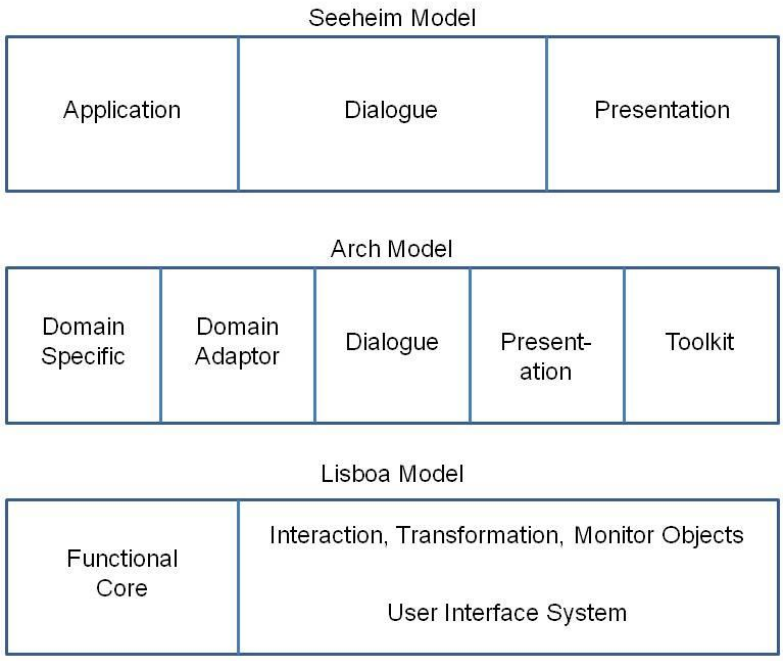


Fig. 4.7 Conceptual Architectural Models [P99]

The framework system architecture maps out the service oriented architectural solution to the problem of encoding heterogeneous metadata as structured XML expressed in terms of java objects and interfaces [EG95]. As a Java based solution the framework architecture follows the model view controller architectural model classified as subsequent high level models to the basic models described by Paterno [P99] as being relevant for such systems beyond interaction. The model view controller and its main adversary the presentation abstraction model are summarized below

- Model View Controller (MVC) Model

  This architectural pattern common to XML and Java subscribes to the shared notion already introduced in chapter 2 pertaining to the separation of the data or domain logic from the user interface constituting the input and presentation [P99] [WP99]. As such the multitier architecture for java applications introduced in the preceding sections and hence the overall architecture of the metadata creation framework resembles the Model View Controller pattern.

  The separation of persistency, container and the client interfaces illustrate the model managing the business rules and data, the controller managing bean transactions and the view as the client interface for interactions and presentation purposes [P99].

- Presentation Abstraction Control (PAC)Model

  Paterno et al. [P99] describes this architectural pattern developed by Coutaz as a three tier interaction pattern similar to the model view controller pattern illustrated in the previous subsection however the middle tier is occupied by control component. The Abstraction component represents the core functionality implementing the processing and retrieval functions in media independent manner hence *"there is an abstract description of the objects to provide to the users"* [P99]. The Control component assumes an intermediary role between the abstraction component and the presentation component managing communication between the perspectives of and linking the other two components. It also *"remembers a local state for supporting multithread dialogue, and maintains relationships with other agents"* [P99]. Finally the Presentation component is concerned with the user interaction and *"perceivable behaviour"* formatting the visual display and presentation of data in addition to receiving user inputs.

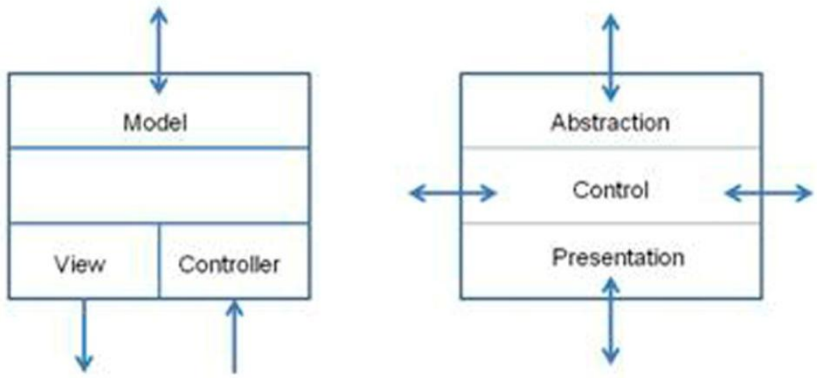


Fig. 4.8 Model View Controller and Presentation Abstraction Controller Architectural Patterns

An implementation of the concepts summarized above presents us with the overall system architecture for a graphical user interface supporting the creation of XML metadata for digital archives.

The implementation reflects upon the functional objectives described in chapter 3 together with their associated binding methods together with related functionalities integrating the aforementioned methods into the metadata framework. The architecture is service oriented consisting of the following tiers:

- user interface tier

  The graphical user interface architectural tier belongs to the principle notions central to this dissertation. This tier provides the users with access to the metadata creation frameworks functionalities with the help of visual perception processes inclined with associated human perceptions [SD71]. The tier resembles the view section of the model view controller pattern [P99] and is implemented with the help of the Java Swing packages for developing graphical user interfaces.

- records management tier

  The collection and management of records in general and in this case bibliographic records in particular, defines according to Wright et al. [RW06] state-of-the-art digital archiving business modeling. Hence this architectural tier is

responsible for the collection and management of the descriptive information i.e. metadata. The implementation is via Java middleware with the help of enterprise beans and respective business models and methods. It is also responsible for managing persistency issues relating to the storage of the relational tag descriptions.

- semantic binding tier

  The marked-up structuring of the descriptive metadata culminates in an XML document validated against a standardized schema. It is this document which is the basis for interoperability within the framework of OAI-PMH and in line with archival preservation standards. The definition of preservation is discussed in chapter 2 runs along the lines of structuring descriptive element tags relating archival context, content and description resulting in metadata heterogeneity. This tier generates heterogeneous metadata structured into XML documents validated against provided schema on the basis of descriptive elements stored as relational elements of archival entities in the frameworks persistence i.e. database.

- integration tier

  The management of users, roles and access to the framework is implemented at this level.

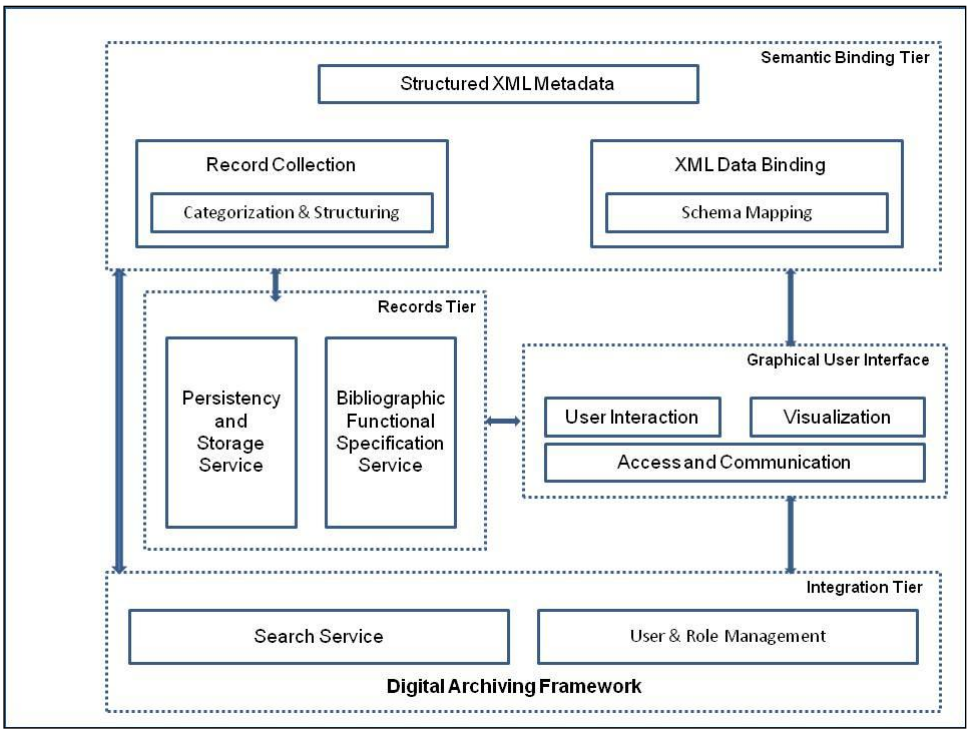


Fig.4.9 Architecture of the XML Metadata Creation Framework for digital archives

The Fig. 4.9 above summarizes the architecture of the metadata creation framework illustrated across the three tiers described in the preceding subsection. Here, the semantic binding tier, the records tier and adjacent graphical user interface together with the integration tier visualize the implementation of the metadata creation framework. The semantic tier conforms to the notion of a semiotic system discussed in chapter 1.2 addressing structured XML metadata encoding and integrating the archival record collection structures with the extensible markup standards. The second tier addresses the data processing aspect of the collected records. This includes structuring along the bibliographic functional requirements, persistence and the interaction with the target users via the graphical user interface. The integration tier deals with the perspectives of the system as a whole and the services offered including selected functionalities such as search and user roles and their management. The subsequent chapter 4.3 elaborates upon this digital archive framework; the selection of implementation classes and their visualisation reflect the architecture described by Fig. 4.9. The framework together with the XML-Binding infrastructure solves the digital archiving problem described in chapter 1.2 and contributes towards a technical consolidation of the digital archive metadata creation process and the management of multi-standard encoding.

4.3 Digital Archive Framework

The majority of research projects in digital archiving focus on the purpose oriented description vocabularies within the framework of existing notions separating content and structure as well as bibliography, text and presentation in addition to their suitability with respect to existing interoperability infrastructure. We claim that this focus constitutes a constraint for XML based structured archiving as it focuses on the technically-versed archivist fluent in XML syntax leading to a neglecting of structuring activities by the majority of archivists mostly from the field of humanities. As a result several technical and interoperability infrastructure exists and remains underutilized. Our research aims to promote structured archiving relieving archivists of the greater part of technical encoding and hand generated XML. This follows the XML principle of the separation of the user interface from the data in combination with the concept of data objects popular among standard programming languages. Such an infrastructure is followed by a separation of tasks in digital archiving.

Metadata recording and collection on the archivist side and adjacent structuring and preservation infrastructure on the computing side. As such computer scientists may access established programming languages and databases technologies developing encoding interfaces for archivists and their data structure and interchange requirements including infrastructures for hierarchical XML mappings. In so doing a greater number of archivists can start to or continue contribute *“state of the art“* encoded documents towards their open archives community, whilst avoiding hard coding. In return, the *marshalling* and *unmarshalling* facilities enable the reading in of new vocabularies implementation of novel XML schema upon existing metadata and records. The general principle is already common to electronic environments already finding resonance in content management outside the semantic archive description spectrum. Our assessment of the aforementioned research objectives by way of a formative evaluation of a selection of scholarly archive projects supports the research question, illustrating the need for usable “embedded” tagging for archive structuring and content management.

4.3.1 Graphical User Interface Classes

A description of the focus of this work as already mentioned in the previous centres around the structuring of descriptive texts and associating them with related text images. Therefore the central components of the metadata framework's application are heterogeneous in nature serving the collection and presentation of images of text as well as the respective descriptive texts themselves. To this order I developed a record creation package *digiarchiv* for collecting descriptions tags in text form and respectively uploading images of digitized archival entities in graphic form. The package structure is aligned to the model view controller architecture whereby the viewer classes are tasked with the user interaction and presentation of digitized material. To this effect image classes are necessary to handle the upload and presentation of the text images and these have been incorporated into an *ImageViewer* component. In Java, images in general are represented by the *java.awt.Image* class encoded in GIF and JPEG formats and presented via the *JPanel* or *JLabel* swing classes. In the *digiarchiv* package the *ImagePanel* component makes use of this class building upon the swing class *JPanel* whilst assuming responsibility for the illustration of the digitized text images. The *ImagePanel* class is then used for presentation purposes as an instance of the viewer class *ImageViewer* which enables the uploading of the digitized graphics. The *ImageViewer* in turn, draws upon the Swing *JFrame* components to variably illustrate and fit the chosen graphic into a graphical user interface delivering the uploaded text image whilst providing for interaction with the user to that effect.

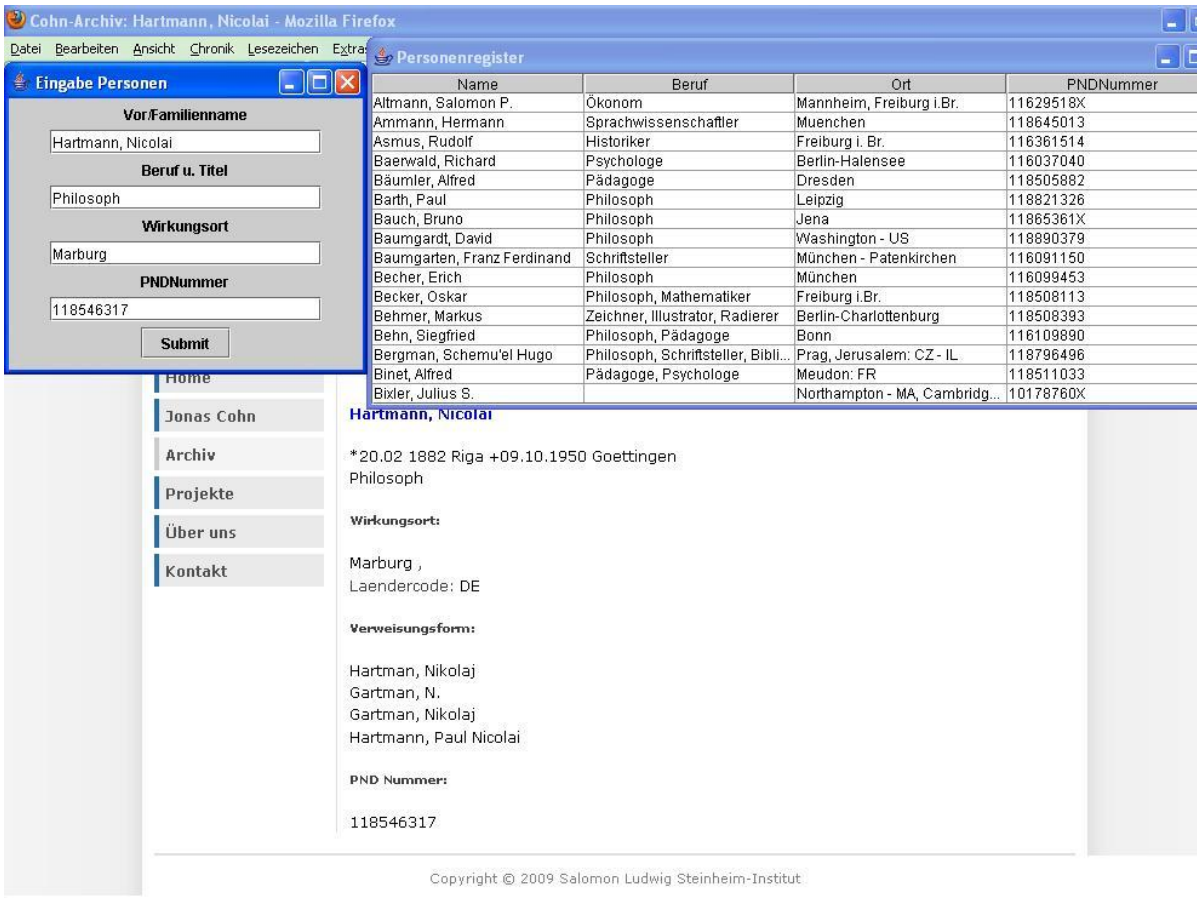


Fig 4.10: Graphical Interface for Person Metadata

4.3.2 The Input Frame Classes

In addition to the image collecting tasks and the respective graphical user interaction based classes, the user is also tasked with collecting the actual metadata classified according to description classes in relation with the entity at hand. To tackle this challenge I have developed simple user input frames based on java swing classes. As illustrated in the figure above, the classified metadata is relational and is reflected upon in the persistence by a relational table whose column definitions may resemble descriptive XML tag elements. The transformation of these column descriptions into XML tags is the subject of XML binding framework and will be elaborated upon in chapter 5 on implementation. However, one aspect of the relational tables and the graphical user interface to note is the association between these tables and user habits elicited as results of the formative evaluation. In this evaluation it was noted that some archivists opted to misuse spreadsheets as databases as they could manually record relational descriptive

metadata as contents of spreadsheet rows with possible description tags as columns. This phenomenon is also elaborated upon in chapter 6 as part of the characteristics of test users. To come back to the swing input frame classes, the implementation is via the *EingabeJFrame* class belonging to the framework package *digiarchiv*. The class makes use of the *JFrame* super class to define the labels and text fields via through which the texts are to be read in. Whereas the *JLabel* and *JTextField* provide for the data entry interface, the association with the relational table is via the *TableModelJFrame* method whose *ActionListener* attributes serve to add texts from the text fields into the table. The table instances the static *TableModelFrame* class hence defining the table structure whilst the *JScrollPane* swing class is used to illustrate the table captions.

Fig. 4.11: Entity Record Collection Interface Person

4.3.3 Descriptive Entity Bean Classes

In chapter 2 we defined metadata in digital archives as virtually accessible texts describing real life artefacts contained and stored in an archive now being presented in digital form. These artefacts were then classified in chapter 3 as bibliographic functional requirements entities which can be illustrated with the help of XML tags. In line with the framework architecture these entities have to be initially collected and stored in relational persistency tables before being reassembled to generate the structured XML documents where they are represented as tagged entities of description.

In order to achieve this, the architecture as described in section 4.1 uses enterprise java beans to represent associations of the archival artefacts with the relevant persistent descriptive data which constitutes our digital archive metadata. To this effect, I have chosen to prototypically represent the functional entities of a digital archive with the help of container managed enterprise beans.

Whilst on the one hand the entity beans allow the implementation of client-server architecture allowing access to the instance by numerous clients, the container managed beans bonds the data represents with the persistency. In alignment with figure 3.1.3 the functional entities Person, Work, Corporate Body resemble a functional triple relating the creator of any archived intellectual artifact, the artifact itself and the archiving entity as provenance i.e. preservation data. With Person and Corporate Body forming a unified ensemble it would be sufficient to summarize a Person and Work bean as it is assumed that framework is for metadata creation in the archive in possession of the work. The resultant prototype then entails a *Person* bean for creators of work and references to persons as content or in context. The root entity represented by a *Work* bean covers all intellectual artifacts associated with a *Person* entity.

- Person
    - Name
    - Occupation
    - Country Code
    - Date of Birth
    - Place of Birth
    - Date of Death
    - Place of Death
    - Also Known As
- Work
    - Title
    - Type
    - Published date

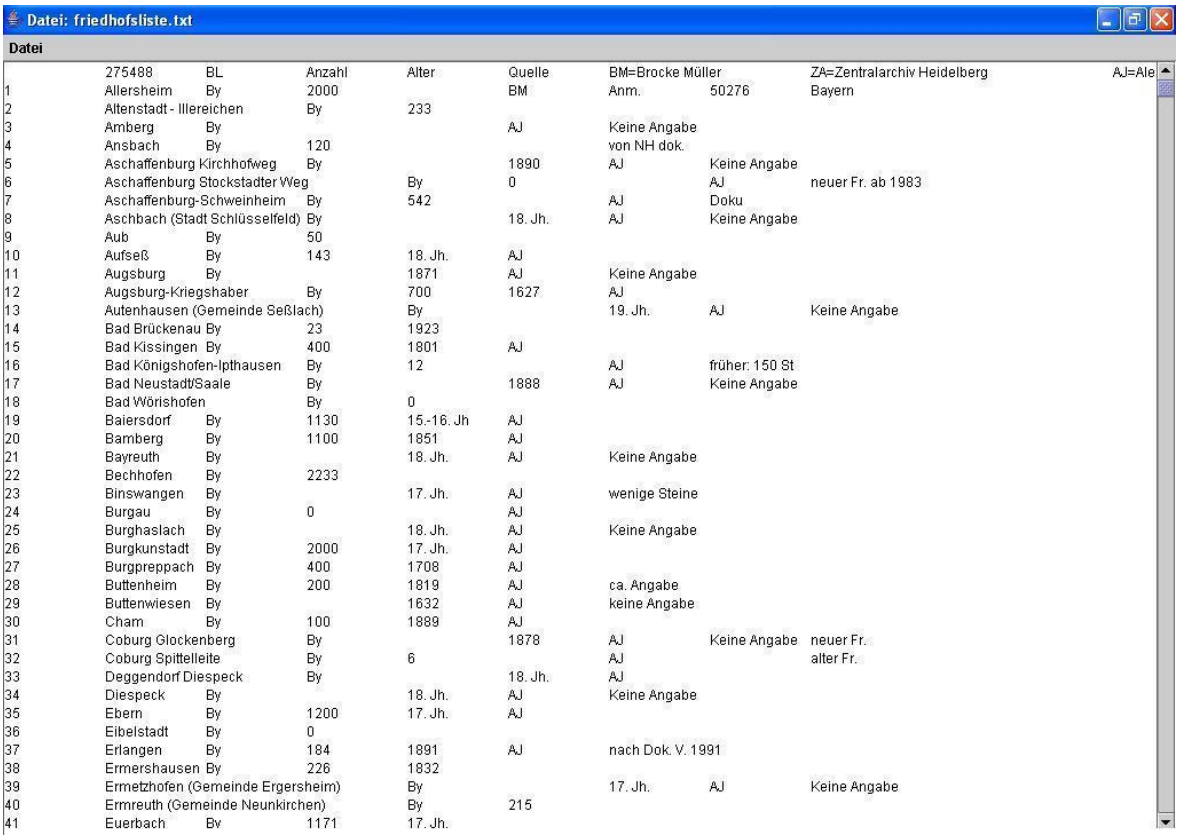
4.3.4 Record Creation Client Classes

In the case of the creation of descriptive metadata, the framework utilizes the enterprise bean multi-tier architecture described in the previous section to facilitate the creation of the descriptive texts with the help of rich clients. The collection and the persistency of the descriptive and preservation metadata remain schema independent. The content of the metadata is to be stored as relational data in the persistency of choice, in this case of the prototype a JBOSS database. Whereas the structuring takes place in the semantic tier, the interaction with the user for record collection purposes draws as in the object classes upon the swing components to present the user with framed text fields for data entry. To this effect, the *digiarchiv* package is extended to include a *DigiArchivSwingClient* component dependent on the *JTextField*, *JPanel* and *JFrame* components of the swing package. Looking at the previous section enlightens us on the *JPanel* and *JFrame*'s role as facilitators for dialogs between users and the functionalities of the framework. The complementing *JTextField* class on the other hand facilitates text editing, allowing us to manipulate texts in the application before storage in the persistency. It is this class which handles our descriptive and preservation metadata texts before they are stored in classified relational tables. The *DigiArchivSwingClient* updates or enriches the chosen persistency tables via the entity bean describing the bibliographic product whose metadata we are collecting.

Summary

The digital archive framework is generally represented by the client server environment visualized by the graphical user interfaces for interaction with the user. The general XML metadata creation purpose has been reduced to a record collection level ignorant to the underlying XML structuring and the resultant digital preservation. As such the framework itself resembles a set of standard interfaces and menus common to the user and familiar to the interaction environment collecting relational description for storage in the persistency. The XML document creation and hence structuring process embedded with the content object structure of the framework and invoke by the binding schema. The user interfaces and swing classes defined for the framework and illustrated above identify the framework user's role as a producer in compliance with the OAIS recommendation's archival lifecycle.

As such the digital archiving framework resembles a record collection infrastructure hosting archival semantic heterogeneity and XML generation whilst interacting with the mostly “luddite” archive user to generate and structure a state-of-the-art digital archive.


Datei: friedhofsliste.txt

Datei

	275488	BL	Anzahl	Alter	Quelle	BM=Brocke Müller		ZA=Zentralarchiv Heidelberg	AJ=Ale
1	Allersheim	By	2000		BM	Anm.	50276	Bayern	
2	Altenstadt - Illereichen		By	233					
3	Amberg	By			AJ	Keine Angabe			
4	Ansbach	By	120			von NH dok.			
5	Aschaffenburg Kirchhofweg		By		1890	AJ	Keine Angabe		
6	Aschaffenburg Stockstadter Weg			By	0		AJ	neuer Fr. ab 1983	
7	Aschaffenburg-Schweinheim		By	542		AJ	Doku		
8	Aschbach (Stadt Schlüsselfeld)		By		18. Jh.	AJ	Keine Angabe		
9	Aub	By	50						
10	Aufseß	By	143	18. Jh.	AJ				
11	Augsburg	By		1871	AJ	Keine Angabe			
12	Augsburg-Kriegshaber		By	700	1627	AJ			
13	Autenhausen (Gemeinde Seßlach)			By		19. Jh.	AJ	Keine Angabe	
14	Bad Brückenau	By	23	1923					
15	Bad Kissingen	By	400	1801	AJ				
16	Bad Königshofen-Ipthausen		By	12		AJ	früher: 150 St		
17	Bad Neustadt/Saale		By		1888	AJ	Keine Angabe		
18	Bad Wörishofen		By	0					
19	Baiersdorf	By	1130	15.-16. Jh	AJ				
20	Bamberg	By	1100	1851	AJ				
21	Bayreuth	By		18. Jh.	AJ	Keine Angabe			
22	Bechhofen	By	2233						
23	Binswangen	By		17. Jh.	AJ	wenige Steine			
24	Burgau	By	0		AJ				
25	Burghaslach	By		18. Jh.	AJ	Keine Angabe			
26	Burgkunstadt	By	2000	17. Jh.	AJ				
27	Burgpreppach	By	400	1708	AJ				
28	Buttenheim	By	200	1819	AJ	ca. Angabe			
29	Buttenwiesen	By		1632	AJ	keine Angabe			
30	Cham	By	100	1889	AJ				
31	Coburg Glockenberg		By		1878	AJ	Keine Angabe	neuer Fr.	
32	Coburg Spittelleite		By	6		AJ		alter Fr.	
33	Deggendorf Diespeck		By		18. Jh.	AJ			
34	Diespeck	By		18. Jh.	AJ	Keine Angabe			
35	Ebern	By	1200	17. Jh.	AJ				
36	Eibelstadt	By	0						
37	Erlangen	By	184	1891	AJ	nach Dok. V. 1991			
38	Ermershausen	By	226	1832					
39	Ermetzhofen (Gemeinde Ergersheim)			By		17. Jh.	AJ	Keine Angabe	
40	Ermreuth (Gemeinde Neunkirchen)			By	215				
41	Euerbach	By	1171	17. Jh.					

Fig. 4.12 Structured lists mapped to text document

Figure 4.12 above illustrates a text document generated from metadata created using the conceptual framework as part of the framework evaluation. Despite the text format the test user recognized illustrated structuring necessary for data interchange and data interpretation as information. The structure also illustrates the scholarly view of the metadata creation process independent of bibliographic and archival data structuring restrictions. A selection of archive use cases is further described in the subsequent chapter 5.

5 Archive Use Cases

The implementation of our metadata creation facility targets archival facilities foremost from the humanities field of study, as to be illustrated in the evaluation described in chapter 6. The framework itself follows an n-tier architectural framework with a role oriented user interface meant for scholarly archives maintained as primary information sources for research and bibliographic interchange. As such this chapter outlines working examples of such scholarly archives and a description of selected framework components. The working examples and the descriptions resemble preparatory aspects towards the evaluation and usability testing of the metadata creation framework to be described in the subsequent chapter. The catalyst behind the digital archiving framework is the digitization project of the Jonas Cohn Archive at Steinheim-Institut. Naturally this archiving project characterizes the proposed framework concept highlighting the key problems of metadata heterogeneity associated with digital archiving. The digitization of the Jonas Cohn Archive faced the challenge of being required to conform to a spectrum of heterogeneous metadata standards spanning across library, archival and internet resource description standards to object and text description oriented markup schema. The metadata creation framework serves to tackle these challenges. The framework resembles a prototypical concept supported by a graphical user interface as an abstraction layer enhancing and tailoring usability to the defined target user community.

The target user community in focus has an archival background in the humanities and is not primarily concerned with the data management aspects of the digitization process however, acknowledging the necessity of standards to which the heterogeneity of the metadata is attributed. The motivating factors lie in the opportunities provided by participation in open archive activities as well as standardization regulations and recommendations of centralized established institution such as the RNA [WK10]. Any archive exposed to the aforementioned factors is subject to upholding the standardization and interoperability requirements dictated by each activity or association. Hence the acceptance, use and proliferation of the metadata creation framework rely on its usability and the navigational assistance of the graphical user interface to the humanities archivist. The latter guides the user’s structured metadata creation as part of the digital preservation

and conservation process or an open archives environment. The usability study resembles the nucleus of my research work providing for an empirical qualification of the acceptance and necessity of the graphical user interface framework.

5.1 The Jonas Cohn Archive

The subject metadata creation framework developed for this dissertation focuses on the notion of a minimized user memory interface. This notion makes use of the standard user interface concept to implement an interaction surface requiring the least previous knowledge or further learning from the target user. To this end the metadata creation framework implemented java based swing graphical interfaces for the menus and interaction with the participating archivists. The java swing classes were chosen due to the programming languages imminent presence in most modern devices commencing with the standard workstation and other every automated machines. As such the user already has a cognitive knowledge of the menus and the file system presentations now being presented to him within the context of digital archiving. Furthermore, the fact that java swing classes are run on several enterprise infrastructure independent of the domain leaves to assume that the user will not be irritated by the presence of such technologies in their employment domain.

The Handwritten Documents

The set of handwritten research journals and correspondence indefinitely on loan from Professor Dieter-Jürgen Löwisch to Steinheim-Institut constitute the foundations of the Jonas Cohn Archive. As a private scholarly archive within an academic institution, the physical archive, its structuring and the classification of the documents and their content are not governed by any regulatory institution or regulations. However, some of the documents and artifacts preserved by the archive are subject to copy, publication and reproduction rights, in some cases by virtue of inheritance.

The correspondence and letters exchanged between Jonas Cohn and other researchers of his time as well as photographic images taken by persons outside the Cohn family in particular make up a large section of these restricted documents. However, Jonas Cohn himself outlined the purpose of his documentation and his research activities along the

notion of open access to information for academia and dedicated this work to future generations after his own.

The archive and Steinheim-Institut in general have an established reputation as reliable primary sources for German-Jewish history and in the case of the Jonas Cohn Archive, New Kantian Philosophy. As reflected by the implicit document type descriptions above, the archival artefacts contained in the Cohn Archive can be characterized according to their origin as handwritten, printed or photographic. In this section, we will look primarily at the handwritten artefacts and their classification where applicable, by their author. By so doing, we expose the archival structure and structural classification of the documents into the following categories:

- Journals
    - Research Manuscript
    - Travel Journal
- Varia (Manuscript)
- Correspondence

Journals

The twenty three handwritten journals reflect the characteristics of a diary and are generally titled as research manuscripts or as travel journals with philosophical discourse as the nucleus of all activities described. Dated between 1911 and 1947, the author classifies them according to his own systematic approach in memories, systematic writings and travel journals. Whereas the research manuscripts are crowned by a working title in Latin, the travel journals enjoy descriptions according to the travel destination giving little insight or indirect reference to the academic content and relevance.

Varia

The Varia on the hand enjoy a direct reference to an academic subject and are written as a collection of thoughts independent of the time frame and the context of the date on which they were started, revised or completed.

Correspondence

Spanning the years 1893-1947 the collection of letters written respectively to and by Jonas Cohn is preserved as part of the correspondence section of the archive. With a cross-section of 260 authors approximately 1100 letters from notable German scholars and academics from the turn of the century, the archive resembles a primary source for research in philosophy, psychology, and pedagogy not only within the German-Jewish context.

Personal Documents

In addition to the academic documentation, the archive contains further private documents of historical importance within the context of Jonas Cohn's role as an academic of Jewish origin at the turn of the century. This cross-section of artefacts spans the target audience of the archive beyond the academic sphere to include local historians, genealogists and other interested parties. The documents complemented by personal photographs and official documents such as denaturalization certificates, exile, re-naturalization and the general context reflecting a lifetime spanning both world wars. The archive seeks the preservation of its artefacts for future generations and for access to the academic world. To optimize this goal in face of the physical decomposition of the mostly paper-based artefacts the archive has taken a series of activities to structure, preserve and enable access.

Problems and Challenges

Leveraging the current structure of the physical archive together with the archive goals and Jonas Cohn's aim of enabling access to information against the information dissemination structures twenty first century reflects the challenges faced by the archive. Whereas on the one hand archival artefacts in particular the handwritten manuscripts succumb to acidic decay and physical disintegration, academic access to archival sources is focusing more on digital information access. In light of these challenges the archive initiated a digitization project aiming to document and preserve archival contents in line with state of the art preservation techniques. Whilst, the chemical restoration of the artefacts made neither economic sense and reflected no relevance to the available

resources, the digital variation provided a scalable obsolescent alternative. However, this option confronted the archivists with structural and bibliographic regulatory conditions outside their scope of research. Having mostly handwritten artifacts, automated pattern recognition proved tedious and expensive and with transcriptions by hand being factually impossible considering the number of artefacts and available funding. Consequently, the archive's metadata and hence their production and structuring are influenced by the aforementioned conditions and the web-based bibliographic environment in which the information dissemination is to take place. The associated problems and challenges may be summarized as follows:

- Handwritten manuscripts and their structure resemble the character of the archival contents and are to be preserved as visible artefacts whilst enhancing virtual awareness. The practical implications of such a prerequisite sees the need for digital images of the artefacts accompanied by structured metadata of the images associating and relating the individual images to the respective documents and bibliographic reference data. Awareness in the computing sense relates the virtual existence of the artefacts to the physical world of the actual document. In this case the historical nature of the artifact, the age and the era are highlighted by facsimile images. The handwritten artifacts as contained in the Jonas Cohn Archive reflected by the Sütterlin handwriting style point towards the era and the historical context in which the artifacts were created. The challenges associated with the transcription of the handwritten material or the optical character recognition are circumvented using digital images of the text generally referred to as digital text images. The set of metadata for images of text as described above follows the notion of structured image description and are to be created as such. Encoding standards for handwritten manuscripts preserved and presented in such a manner also require presentation structures to present them chronologically in book format.

- Bibliographic records of the archival artifacts and associated summaries of the content represent text and must be structured accordingly. This set of metadata represents the second set of metadata governed by bibliographic regulations however describing the same archival artifacts. Challenges facing the archive include a restructuring of the archival artifacts along bibliographic regulations in preparation for a consolidated metadata repository.

- Extensible markup structured metadata together with associated structuring for presentation, preservation and obsolescence. The archival contents are to be presented via the worldwide web representing the propagation medium and in line with the state-of-the-art semantic and resource description infrastructure. Furthermore, bibliographic resources and the structured marked-up images and records are to be linked within the markup structures, in particular the relation between text images, associated records and text summaries of the content.

- Reorganization of the archival contents along the RNA [WK10] bibliographic recommendations and including the products of digital preservation efforts. This includes catering for records of the digital images, the microfiche and the digital films of the artifacts. The archival content is thereby enriched by the products of the digitization; however these resemble further artifacts and need to be taken into stock. In other words, the digitization results in a whole new set of archival products and product categories requiring semantic and bibliographic markup as well as a relation to the original artifact document.

The problems and challenges described above highlight aspects of the metadata creation process required to assess the usability of the proposed metadata creation framework and evaluate the implementation by means of the graphical user interface with which the archivists interact with the framework. These highlighted aspects represent the framework and the boundaries along which the usability testing and the summative and formative evaluation are based. The following subsection illustrates the resultant structuring implemented alongside the digitization whilst serving as the basis for a comprehensive usability testing of the framework with the Jonas Cohn Archive in the role of the test subject and on the basis of the usability tests to be described in the subsequent chapter 6.

Cohn Archive Digital Edition

The digital edition of the Jonas Cohn Archive is characterized by the RNA [WK10] bibliographic regulations in accordance with the recommendations of the German Research Foundation DFG [DF10]. As such the digital archive resembles a reflection of the physical archive complemented by structured metadata enabling information interchange, long term preservation and obsolescence using state-of-the-art extensible markup. This characterization follows the notion of an object relational classification of

archival artefacts in records, digital objects and facsimile images (i.e. digital image objects) and their respective storage and preservation in digital media. The RNA builds upon the principle of “Personennormdatei” (PND) and the “Gemeinsame Normdatei” (GND) to index authors and institutions in the bibliographic context.

This principle preserves the relation of an author his or her publications and their variations in addition to multiple references of an author via different names or nom de plumes and follows and object oriented association based on the author name as the root object. As such all references to publications implemented along the PND/GND or related recommendations are transmitted as metadata to the centralised national library index for long term preservation purposes. As a result the digital edition of the Jonas Cohn Archive differs in structure and constitution from the paper based physical archive and these differences are reflected by the following characteristics:

- a digital preservation oriented archival concept encompassing information storage
- a heterogeneous object content structure
- a reorganisation of the archive along bibliographic lines
- an object centred archive structure
- an object relational archiving systematic
- a unified medium for archive content management and presentation
- a partly specified target audience

Focus on Digital Content

The characteristic differences in archive structure and focus follow the trail of a shift in focus from paper based preservation towards the notion of digital content. The characteristic properties of this digital content are in turn influenced by the digital objects making up the nucleus of the preservation activities. Subsequently, the focus of any archive management software, its respective usability and the usability testing shifts along this notion of digitization as these influence target user roles and tasks [DR99]. In the case of the heterogeneous metadata creation framework and the associated usability testing, the transition towards the digital archive notion within a bibliographic framework offered the basis for the design of the usability tests and the choice of the test subjects. An analysis of the tasks associated with this transition and the respective resultant archivist activities define according to Dumas and Redish [DR99] the guidelines for a customized

usability test design. The usability heuristics are also influenced by the digital objects i.e. the metadata produced by the framework system as well as those processed by the system within the framework of object relational preservation and presentation.

These objects dictate a further structuring of the content and the metadata according to the artefact type and can be classified as follows:

- Bibliographic data
- Digital Text Images (digital images of text)
- Digital Microfilm
- (Microfiche) meta information and description storage

In the context of the Jonas Cohn Archive, the archival description and the characteristics described and defined above governed the structure and the design of the usability tests which constitute the formative and summative evaluation of the metadata creation framework. The design of the tests along the notion of digitization of heterogeneous descriptions of object variations of the same artefacts provided guidelines for the choice of heuristics and the methodology. The usability tests, together with the recommendations made by the research foundations with regards to integrated metadata creation, the archive business environment and the state of the art propagation technologies helped outline the research question posed by the task and the exposure of the proposed solution.

5.2 The Hegel Archive

Artefacts

The establishment of the Hegel Archive is strongly related to the publication of the collected works of the philosopher Georg Wilhelm Friedrich Hegel (1770-1831). The collection is complemented by Hegel's studies and the works and correspondence of the philosopher and economic reformer Friedrich Heinrich Jacobi (1743-1819). As an institute of the Faculty of Philosophy of the Ruhr-Universität Bochum the focus and structure of the archive is purely governed by academic an academic drive and respective principles. Consequently, the priority awarded to metadata creation and structuring along bibliographic reflects a different operational level within the context of this dissertation

work. As a result the participation of the test users from the Hegel Archive resembled a reflection across the subpopulations representative of the target user groups [JN93]. The archival structure is dominated by the collected works origin and constitutes a mixture of preserved primary and secondary sources. This structure influences the nature of collected metadata and the focal point of the targeted metadata creation process. Similarities to and overlaps with the Cohn Archive on the structural side provide the basis for standardized interfaces and information management concepts.

Hegel Archive enjoys an established reputation as a reliable source for Hegel Philosophy across the board and across the cross-section of authors, scholars and philosophers. The origins of the archive further influence and reflect the artifacts contained by and the character of the Hegel Archive. The characteristics structure describes microfiche, book publications, lectures and correspondence. In this section, we will look at the artifacts and their classification and the object relational approach and the possible structural mappings. The archival structure as such and the structural classification of the documents is defined according to the following categories:

- Collections
    - Lecture Notes
    - Publications
- Studies
    - Volumes
- Manuscripts
    - Writings
- Correspondence

Artefact Document Structure

The artifact document structure of the Hegel Archive is similar to that of the Jonas Cohn Archive, pre-digitization preservation measures resulting in a collection of microfiche and the philosophy research focus complement the structural overlap. The categorization of the artefacts according to the document structures mentioned above and their respective mapping to the categories recommended by the RNA regulations [WK10] align both archives to the same user test population. The differences and organizational attributes of the archives refine their affiliation to the test population qualifying subgroups within the

user test group for the evaluation phase of the subsequent chapter 6. Conforming as a result to the recommendations by Dumas and Redish [DR99] and Nielsens approach to Test User Selection for the discount usability method [JN93]. In other words, the document structure of the Hegel Archive artefacts can be mapped to that of the Cohn Archive and vice versa guaranteeing a structural standardization of the collected metadata and the object relational notion within a bibliographic context. The digital content of both archives is therefore similar in structure and nature up to the point where digital text images represent physical material and their dissemination in digital form as an edition.

Summary

In summary, the Hegel Archive shares structural metadata attributes with the Jonas Cohn Archive however not preceding information description and object relational bibliographic data. Furthermore, the notion of bibliographic indexing on the basis of PND (GND) is implicit and of secondary value as the archive organizational structure is guided purely by philosophical research. This in turn influences the target user subgroup and the usability of the graphical user interface due to the broader user group and the unavailability of a dedicated archivist for metadata creation. The heuristics of the discount usability approach [JN93] are meant to deal with these challenges within the framework of the client server architecture.

5.3 Planning the Evaluation

The archive summarized above represent a subsection of the population of target user groups selected for the evaluation process to be described in the subsequent chapter 6. The short description of the archives and their metadata structures where applicable or aspects influencing the metadata structure helped design and structure the evaluation process in general and the usability tests in particular. The entire evaluation process was governed by necessity to develop usable infrastructure for the target group as summarized by the usability principles outlined by Dumas and Redish [DR99] as follows:

- *“Usability must concern any group developing any product that people are going to use”*
- *“Usability has to be thought about, planned and designed”*

Together with Nielsen's reference to the difference between the notions of utility and usability as in the access to functionality provided by the system [JN93], the planning and design of usability and usability testing must optimized along the goals behind the evaluation. In the case of the metadata creation framework, the following principles

- Optimize framework usability
- Test users must be representative of real users
- Test users must do real tasks
- Evaluator must observe and record test users' actions and what they say
- Analyze the results, diagnose real problems and recommend solutions

Applying these principles to the test user population represented by the Jonas Cohn Archive and the Hegel Archive and their respective archival structures as described above laid the foundation for a solid and representative evaluation on the basis of the usability discount method. As such the generalized goals for the usability testing include:

- Graphical User Interface acceptance
- Testing whether user navigation for simplicity
- Testing acceptance of virtual archive as a replacement for the physical archive
- Testing user response to bibliographic recommendations
- Testing user awareness towards structured archiving approach
- Velocity of the user transition into productivity
- Testing user acceptance of dedicated tools within a homogeneous medium
- Testing user awareness of metadata heterogeneity and possibility of implementing publication workflows across presentation media

The archive use cases described above represent the test user population of the valuation phase to be described in the subsequent chapter 6. The focus of the use descriptions targeted the characterization of the test user population and their properties. Figure 5.1 below illustrates a typical XML fragment containing and describing archival metadata as seen by the metadata creation framework. The figure serves to illustrate the complexity of the markup structures whilst stressing the level of usability needed to accommodate users from the described target test population. The principle behind the graphical user framework encapsulates the markup encoding structures embedding them into abstract windows is elaborated in the subsequent subchapter.

```
<mets:mets xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
    xmlns:xlink=http://www.w3.org/1999/xlink  xlmns:mets=http://www.loc.gov/METS/
       xsi:schemaLocation=http://www.loc.gov/METS/ http://www.loc.gov/mets/mets.xsd>
    <mets:dmdSec ID="md92018">
      <mets:mWrap MIMETYPE="text/xml" MDTYPE="MODS">
        <mets:xmlData>
            <mods xmlns=http://www.loc.gov/mods/v3
         version="3.0"
         xsischemaLocation=http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-0.xsd>
            <titleInfo>
                <title> Pandocheion heteron (deuteron) </title>
           </titleInfo>
           <name><displayForm>Cohn, Jonas </displayForm></name>
         <originInfo>
         <place>
            <placeTerm type="text"> Birmingham </placeTerm>
         </place>
            <dateIssued> 27.01.1946</dateIssued>
         </originInfo>   </mods>
  <mets:xmlData>
</mets:mdWrap></mets:dmdSec>
```

Fig:5.1 XML Code Fragment

5.4 Abstract Window Principle

State-of-the-art digital archiving toolkits aim to simplify the archiving business and its associated processes by providing automated functionalities incorporated into graphical user interfaces making it easier for digital archivists to structure and manage their archive contents. As technology improves and structuring standards change these graphical interfaces will become the gateway for archivists taking part in the digitization of the archives and their data management processes.

Open Archiving and inherent interoperability represented by data interchange now more or less dictate the management processes and architectures involved in digital archiving. These dictates can now be incorporated into independent graphical interfaces sparing the user the effort of learning and manually encoding novice XML standards. The abstract windowed encoding principle combines the data abstraction into content object classes which can be represented by any modeling language and the embedding of XML code within a data binding framework.

In addition to providing for user interaction and communication interfaces, digital archiving frameworks prove to be efficient intermediary elements between data persistency technologies and the structured XML encoding. To this respect XML generation and transformation interfaces serve to mediate between the two aspects with technological novelty incorporated in the widely accepted and platform independent java programming suite.

Summary

The use cases described above represent the typical academic archives and the associated structural adjustments associated with the digitization of archives of similar characteristics. The scenario of academic staff managing primary research sources on the basis of bibliographic preservation infrastructure exposes the challenges typically faced by the archives. The use cases on the other helped the design and structuring of the usability tests and the assessment of the acceptance of the proposed framework by the target community. The following chapter 6 describes the evaluation stages of the dissertation work commencing with a formative evaluation of the entire test population in preparation of the summative evaluation of the framework system prototype.

6 Evaluation

Digital information systems have become increasing popular alongside numerous computer based supportive tools and the internet, the classical symbol signifying the digital age. As popularity increases so do questions regarding the necessity of not only of digitization but also the multitude of supportive tools. Possible answers to these questions can be provided by an evaluation of the software tools in question based upon chosen relevant criteria. Given the novelty of digital archive management systems tool an evaluation is an imperative aspect for the assessing the tool's usefulness, effectiveness and its success. This chapter will therefore discuss the goals, criteria as well as methods and the results of an evaluation of the archive management application.

Evaluation is defined in the Cambridge dictionary as the process of judging or calculating "the quality, importance, amount, or value of something". Wilson [TW85] complements this definition to include the quantitative measurement of criteria which "indicate when an objective has been met", bringing in, in effect the scientific nature of evaluation as a field of research in modern information systems. In other words, evaluation in digital information systems represents the analysis of the system based on a selection of quantitative criteria resulting in a qualitative measure of the system's effectiveness and success. Classification of evaluation is based on the purposes and implementation and can be subdivided into two main categories [FS09] [LMB02]:

- **FormativeEvaluation**
  Implemented parallel to the software development process serving to expose conceptual weaknesses and helping to improve products and programs by providing information within the planning and development stages. The evaluation process can be likened to internal quality control as implied by Lockee et al. [LMB02] ensuring "quality in a unit or course before release." It involves mainly small user groups guided by flexible evaluation methods and is suitable for evaluating instructional design as well as interface design issues, the latter of which is applicable to this dissertation or cultural heritage systems. In other words, a formative evaluation gives us information on navigation, aesthetic etc.

- **SummativeEvaluation**
  Serves to measure the effectiveness, efficiency and usability of a system, in other words whether or not the system achieves the intended goal or not and is hence implemented upon completion of the development process. Lockee et al. liken summative evaluation to a food critic “reflects how well the final object works in the real world” [LMB02]. The measurement units may be user acceptance, efficiency and knowledge transfer providing for a measure of the achievements from a qualitative and effectiveness point of view [FS09].

In addition, a summary of the definition and classification of evaluation identifies evaluation as a powerful tool necessary for the development of effective usable digital information systems. However it also highlights the need to identify and specify the units of measurement and evaluation methodology tailored to the purpose and audience of the system being developed. Wilson [TW85] lists the general evaluation criteria as:

- Success
- Efficiency
- Effectiveness
- Benefits
- Costs

Evaluation Criteria

The identification of relevant criteria is of great importance as these criteria form the basis for the development of the evaluation measures [TS04] and hence the identification of goals and objectives [TW85]. However, investigating the reason why we have to evaluate the information system is addressed by the notion of a basic criterion [TW85] [TS04]. The notion of *relevance* is widely accepted as the basic criterion for digital libraries and information retrieval systems. Also widely accepted in literature is the acknowledgement that this does not imply to all information systems and especially not to all digital libraries and cultural heritage systems. In light of the fact that evaluation is classified according to the purpose of the software being developed it is only logical that this also applies to the evaluation criteria. Saracevic [TS04] summarizes the “most often used criteria” of which most relevant to this dissertation are as follows:

6.1 Usability

Defined by the ISO as *"the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* in other words *"the quality of interaction between the 'User' and the 'System.'"* As a basic criterion *usability* is considered a *"meta term"* i.e. it doesn't handle data but defines the quality of interaction serves though as a general term for sub criteria considered *"specific criteria"* summarized as follows [TS04] [F07]:

- ***Content*** (of a portal or site)
    - accessibility, availability
    - clarity
    - complexity (organization, structure)
    - informativeness
    - transparency
    - understanding
    - adequacy
    - coverage, overlap
    - quality, accuracy
    - validity, reliability
    - authority

- ***Process*** – task implementation
    - learnability to carry out
    - effort/time to carry out
    - convenience, ease of use
    - lostness (confusion)
    - support for carrying out
    - completion (achievement of task)
    - interpretation difficulty

- sureness in results
- error rate

- ***Format***
  - attractiveness
  - sustaining efforts
  - consistency
  - representation of labels (how well are concepts represented?)
  - communicativeness of messages

- ***Overall assessment***
  - satisfaction
  - success
  - relevance, usefulness of results
  - impact, value
  - quality of experience
  - barriers, irritability
  - preferences
  - learning

Usage / Usefulness

Not considered as a complete evaluation rather as a basic criterion concerning the “User” and “Content” components translated into actual relevance, task relevance as well as type and level of resource relevance and involving studies of [F07] [TS04]:

- usage patterns
- use of materials
- usage statistics
- who uses what, when
- for what reason/decisions

In summary, a determination of evaluation category and criteria collectively trigger the evaluation process planning and help guide developers towards the quantitative measures of value. However, a comprehensive evaluation method is imperative for a sensible deployment of the identified basic criteria (*relevance*). In the subsequent subsection, we will look at a couple of evaluation models suited to archive management systems and in which usability and usage/usefulness are of relevance.

6.1.1 Discount Usability Engineering

In addition to the structuring of collected metadata, the framework described in this dissertation aimed at simplifying the data collection and the generation of structured XML for the non-technically versed user. To this effect, aspects of the former constitute a prototypical graphical user interface described in chapter 4 and these were tested with the help of usability engineering and related user tests within the framework of the system evaluation. In order to access the suitability of the developed framework for its intended tasks, I have chosen to follow Nielsen's evaluation theories [JN94] for user interaction systems which focus on empirical usability and user experience assessments.

Nielsen's theories summarize usability and user oriented evaluation under the umbrella term *Usability Engineering* as being the set of techniques that assess software suitability.

In acknowledgement with these principles Nielsen introduces the *Discount Usability Engineering* [JN94] as a process method accompanying the design of interactive applications and their user interfaces. Despite being considered an expert method illustrating the greater number of challenges faced when designing interactive applications and their interfaces, the discount usability engineering can be minimized to its heuristic evaluation element, which together with the additional techniques listed below constitute the discount engineering method's set of techniques.

- User and task observation

  The discount usability engineering method encourages user involvement at the early stages of the software development in order to be able to feedback eventual results into the development process. User and task observation belong to this type of user involvement and have for the purposes of this dissertation been dealt with as part of the formative evaluation and the framework task analysis discussion in chapter 3. Nonetheless Nielsen's method describes this as discount task

analysis incorporating evaluation rules which are in agreement with the implemented formative analysis summarised by Nielsen in the following description *”observe users, keep quiet, and let the users work as they normally would without interference”* [JN94].

- Scenarios

  Scenarios as a technique aim at extracting frequent user feedback in a cost effective flexible manner implemented in *“simple prototyping environments”* or as *“paper mock-ups”* [JN94]. A further advantage is reflected by the exemption of complex software tools from the prototyping environment in favor more simplified programming environments which are in turn easier to learn. The technique belongs to the family of horizontal prototyping which in principle require a trim down in complexity by eliminating sections of the full system and functionality levels. Whereas this results in a user interface layer, vertical prototyping on the other hand implements the full functionality of a reduced selection of features resulting in a fully functional subsection of the system in question [JN94]. Horizontal prototyping of the metadata creation framework is reflected upon by the test design of the formative evaluation implemented in chapter 3. In this case, the functionality of the metadata collection facility reduced in complexity illustrated the record facility by the web-tier as a web-based interface layer. For the Summative evaluation vertical prototyping scenarios with developed swing interfaces take their place and are described together with chosen tasks in the subsequent implementation subsection of this chapter on evaluation.

- Simplified thinking aloud

  Although this technique sounds more like a psychological analysis than a software evaluation technique, it is equally popular as a novel user interface assessment technique among usability and user interface experts [JN94]. The technique sees individual test users completing a selected set of given tasks whilst *“thinking out loud”*. An analysis of the verbal comments gives an insight not only into *what* the interface endures but also *why*. Whilst traditional thinking aloud requires the filming of test users and a detailed protocol analysis, simplified thinking aloud sees this method as being intimidating and less effective in comparison to cost and training expenditure. Instead having evaluators take notes to protocol the experiments provides a more solid basis for an effective data analysis in place of video

recording. Nielsen quotes a survey of 11 software engineers whose assessment sees simplified tests as being *"almost twice as useful as video protocols"* [JN94].

- Heuristic evaluation

  Similar to "simplified thinking aloud" heuristic evaluation sees a minimization of evaluation test criteria found intimidating by developers to a set of 10 heuristic rules. On the one hand, this targets a reduction in complexity by "two orders of magnitude" [JN94] and on the other hand focuses the summative evaluation on the basic usability principles listed below and the usability heuristics principles. The evaluation is done by having expert users assess the interface and provide opinions on its positive and negative aspects based on the heuristics.

6.2 Evaluation Aims

In simple terms, the main aim of this summative evaluation is to assess the overall acceptability of the developed metadata creation framework. This overall acceptability is derived from the system's social and practical acceptability with the latter focusing on aspects of cost, compatibility, reliability and usefulness. Whilst cost related factors play no role for work in this dissertation as it is part of an academic research, the usefulness of the framework application as a tool assisting digital archiving in academic archival and edition projects is of importance. As such the evaluation aims to assess this assistance in terms of usefulness and against the set of heuristics specified by the discount usability engineering method. The reliability and validity of the summative evaluation plays a major role as the framework and its specification serve a small section of archive types mainly found within the academic environment. In contrast to the formative evaluation, the summative evaluation sums up the development process whilst assessing the overall quality of the developed interfaces in light of possibilities of web and swing based clients for metadata collection. As the target users are more interested in the data collection process with the structuring and conforming to interoperability guidelines taking a minor role, evaluation will not encompass functional testing but be confined to user interaction. The purpose can therefore be summarized by the usability categories listed below which in turn translate into the usability principles of a heuristic evaluation to follow summarized in the subsequent table 6.1.

- Easy to learn

- Efficient to use
- Easy to remember
- Few errors
- Subjectively pleasing

Due to the multidimensional nature of usability, its aims and hence those of the evaluation of its aspects are prone to conflict and therefore require user and task analysis based priority setting. In order to assess such priorities project specific usability metrics expressing the chosen parameters in operationalized measurable terms have to be defined. These operationalized terms are concurrent with the usability principles outlined in table 6.1 below and resemble a metric representation of the usability aims and hence the overall acceptability to be tested as part of the principle aim of the actual evaluation. To this end usability testing and the respective procedural elements illustrate an analysis of the aims and goals of the summative evaluation and the associated empirical analysis'.

Simple and natural dialogue
User language
Minimize user's memory load
Consistency
Clearly marked exits
Feedback
Shortcuts
Good error messages
Prevent errors
Documentation

Table 6.1 user interface usability principles[JN94][DR99]

Table 6.1 outlines contemporary usability and user interaction principles [DR99][JN94] in consistence with common principles for interaction design. The interaction principles

derived from user expectations aim towards optimizing cognitive communication between a system and its users. As outlined in the preceding subsection, the empirical translation of the usability principles into usability heuristics resembles the basis for a qualified analysis of the usability of a system. As such providing the basis for and outlining evaluation aims.

6.3 User Testing

An evaluation of a user oriented software system and in particular one where the user interface belongs to the main elements of interest culminates into usability tests using target users of interest as test subjects. To this end an assessment of the target users and their task related habits becomes an essential aspect of the evaluation inherently encompassed with the act of user testing. Whilst this act belongs to the usability engineering model's lifecycle, it reflects the *"Know the user"* notion propagated by Nielsen et al. [JN94] in the discount usability engineering method as illustrated according to the following sub-stages:

- Know the user

  This methodical analysis familiarizes the system developer with the user subjects and their tasks for the purposes of incorporating user and task characteristics into the development and test processes.

  - Individual user characteristics

    Concerns the classification of the people meant to utilize the system being developed. Classification aspects may include experience, age or educational level or simply the preparedness to sacrifice time and energy in learning how to use a system. Whereas most of these criteria are valid for systems publicly available for open use, the case of a framework to assist digital archiving focuses more on a special interest group. This special interest group is further characterized by the associated reasons behind the motivation for digital archiving. In this special interest group time constraints pose a threat to associated preparedness to learn new systems as archivists are busy with content associated aspects of the archive in the form of edition work digital or otherwise.

  - User's current and desired tasks

This sub-stage has already been dealt with as part of the formative evaluation. It deals with an investigation of the user's current and future tasks and results in a task model serving as feedback for the system and it this case framework development process.

- Functional analysis

The functional analysis extends the assessment of the tasks beyond the simple tasks analysis to include the reasoning behind carrying out a task then referred to as the function. The identification of necessary tasks and their differentiation from the called *"surface procedures"* resembles such an analysis. Nielsen et al. [JN94] associates this with the user approach to a task and sees it as being concurrent to the task analysis.

- User evolution

In the same manner that technology is rapidly changing, this sub-stage acknowledges the exponential development of the target user with respect to sophistication and the manner in which the system can be used in the future. This dialectic phenomenon commonly referred to as the *"co-evolution of tasks and artifacts"* [JN94]. The user practice of using spreadsheets as database, also observed during the formative evaluation of the metadata creation framework provides a lucid example of this phenomenon.

An initial acquaintance with the tasks and users of a digital archiving framework and the resultant task models have been described in chapter 3 and its associated formative evaluation. For this purpose users were visited in their respective environments i.e. onsite visits and an assessment of the user characteristics and the variability in tasks carried out. The results of these visits constitute an overview of the formative analysis and help categorize the characteristics and differences of our target users. The latter can be illustrated as a three dimensional user cube giving guidelines for getting test users representative to the target users and how to effectively model, structure and plan the usability tests.

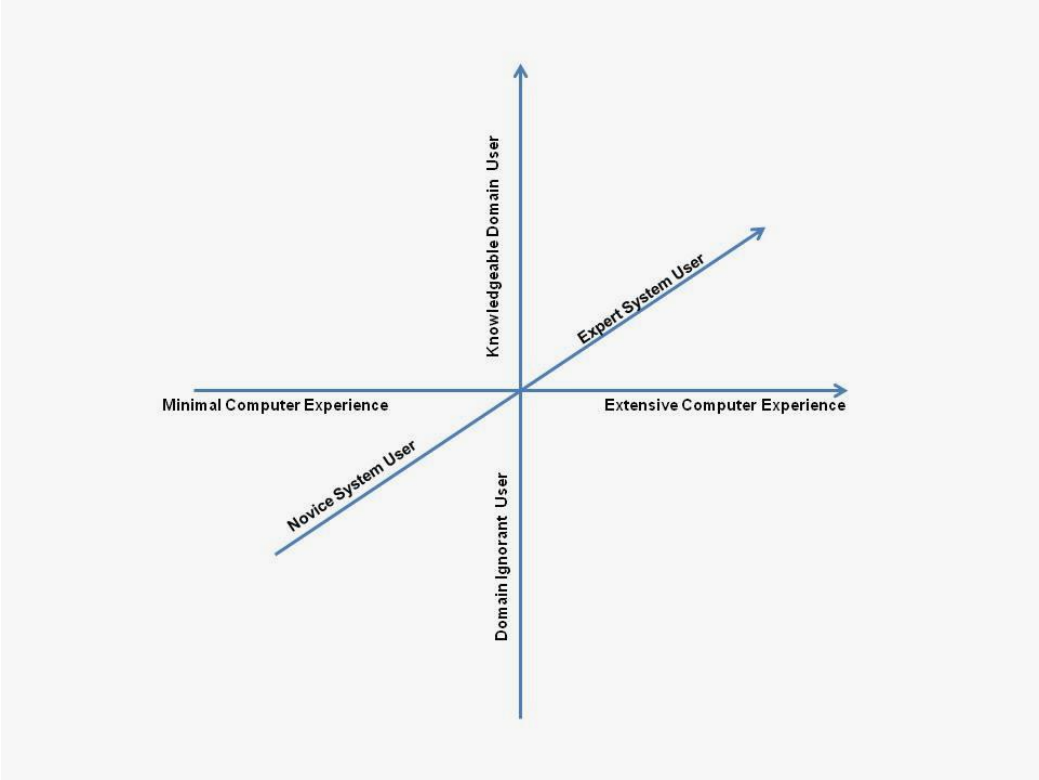


Fig. 6.1 The 3 dimensions of user categories [JN94]

6.3.1 Test User Recruitment

The selection and commitment of test users influences the empirical quality of the summative evaluation and is a major attribute of the evaluation results. In the previous subsection we have discussed the analysis of users and their characterization with the resultant three dimension categories illustrated in the figure above. The test user selection must be oriented to these three dimension categories hence representing the broad spectrum of the special interest target user groups. In addition to these dimensional categories, the chosen test users should be representative of the targeted end users which for the case of digital archiving metadata are scholarly academic archives or respective edition projects. The target users as I have called them up to this point of this chapter may be classified in groups as user types or as part of a general population of users. Special interest software developments prove to be unique as test users may be specified as individual testers as has been the case for the metadata creation framework. Although this simplifies finding and selecting test users, it is often and has been plagued

by user availability difficulties as the users are preoccupied with their primary occupation. Having test institutions select users provided with a leave of absence may distort the dimension of user categories as well as the entire evaluation results should the institutions provide either their best or their least-best users depending on their individual criteria. Selecting users' balances between the usability attribute *"demoability"* and the actual software use.

Nielsen's Sample Size Theory Controversy

The notion of „less is more" postulated by Nielsen's Graph [JN93] described in section 6.3.2 is the subject of academic discourse on usability. The key question centres along the sample sizes for usability evaluation and consequently usability tests. In light of the implementation of Nielsen's Discount Usability theory [JN94] as the basis for evaluating the metadata creation framework described in this dissertation, the key question of usability evaluation sample sizes is also briefly dealt within this subsection. According to Sauro the question of sample sizes in usability testing can be traced back to the problem of "diminishing returns in problem discovery" [SJ10]. Between the years 1981 and 1982 described by Sauro [SJ10] as the Pre-Cambrian Era, the notion supporting the sufficiency of observing five or six usability test users was coupled with a model based on the binomial distribution. The Cambrian Explosion (1990-1994) purportedly a consequence of the widespread use of graphical user interfaces saw a multitude of academic papers supporting the notion however based on earlier research. The key notions debated upon, the first two of which confirmed earlier theories [SJ10] included:

- "Additional subjects are less and less likely to reveal new information"
- "The first 4-5 users find 80% of problems in a usability test"
- "Severe problems are more likely to be detected by the first few users"

The question of the relation between severity and frequency remained. The other two notions constitute the nucleus of Nielsen's [SJ10] summary of the previous decades work and the subsequent Nielsen's Graph [JN93] illustrated in Fig. 6.2. The graph and the theories have accompanied usability testing then on. Despite the purported acceptance of the theories, the description of the graph as the "parabola of optimism" by Sauro [SJ10]

illustrates further debate in opposition to the notions culminating in criticism of the notion's legitimacy.

In general, criticism of Nielsen's Graph and the notion of a sample size of five test users focuses on the binomial distribution based model and its relation to the number of problems discovered and their severity. Whereas problem revelation is adequately dealt with, issues pertaining to problem occurrence and frequency particularly with respect to the sample size model remain controversial. Sauro [SJ10] attributes these discrepancies to the test designs, particularly open-ended tasks and variations in user types. Consequently subsequent debate agrees on the legitimacy of Nielsen's graph however coupled with the test design and parameters as validity boundaries. The coupling and optimization of the binomial model to accommodate variability in problem frequency remain open to research.

The implementation of the discount usability engineering method during the dissertation work took into account the parameters and boundaries in which this method is deemed legitimate and valid. The sample population consists of academic archivists in the humanities seeking digital editions of the archives and the sample size is by nature of the population also limited. As such the graph and the notions postulated by Nielsen were best suited to the dissertation question and served to provide answers pertaining to usability and user acceptance within the framework of the validity parameters. Sauro [SJ10] summarizes the legitimacy of Nielsen et al. theories as follows:

- Revelation of 85% not of ALL of the problems, but of the more obvious problems
- "The sample size formula only applies when you test users from the same population performing the same tasks on the same applications"
- Select a minimum problem frequency you wish to detect as the test parameter hence outlining the chances of detecting problems with "that probability of occurrence".

In other words, Nielsen's postulate is sufficient for the purposes of determining general user consent towards the notion of guided metadata creation with the help of intermediary technology within the homogeneous sample group of archivists in the humanities. The merits of the different statistical methodology behind the notion are secondary weighed

against expected results and practicality. Macefield [MR09] shares this view on the question of sample sizes for usability studies and further gives a practitioner's view whose focus goes beyond problem discovery on interfaces. As such, Macefield acknowledges Nielsen's theory to a study carried out by Nielsen and Virzi however on a "95% confidence level and an error margin of +/-18.5%" [MR09]. Furthermore, Macefield relates problem discovery with context criticality leaving room for validation of discovered problems by the target user groups on the basis of validated criticality measures. This view reflects and applies to the scenario of the metadata creation framework described here. Focus lies on the target user acceptance of an interface assisted metadata creating system and less on the individual interface and associated problems to be discovered. Further parameters including time, budget and test user availability also influenced the choice of the usability method. In other words, the selected usability testing method best provided a practical compromise between effective results and the usability parameters assessed given a specified test population and test scenario. The selected method provided concurrency in view of practical usability testing and the ideal scenarios purported by statistical theories and models.

Table 6.2 below further outlines contemporary testing methods [JN94] for usability testing and user interaction design. The methods illustrate common data collection and enumeration methods in relation to user acceptance, interaction design and user interface development. Whilst not all of the aspects covered in table 6.2 were part and parcel of the evaluation effort of the dissertation, the overlaps between methods and the resultant empirical analysis are best illustrated. In addition to recommended evaluation stages within the development lifecycle, the table guidelines user test sizes leveraged against, the method, stage and associated advantages and disadvantages.

Testing Methods

Method	Lifecycle Stage	No. of Users	Advantage	Disadvantage
Heuristic	Inner cycle of iterative design	None	Identifies usability problems	No real users finds no surprises
Performance Measures	Final testing summative evaluation	>10	Eases empirical comparisons	Doesn't expose individual usability problems
Thinking aloud	Formative evaluation	3 - 5	Pinpoints user misconception. Cheap testing	Unconventional for most user types- luddites and introverts
Observation	Task analysis follow up studies	3 or more	Ecological validity exposes real tasks	Difficulty organizing appointments & test control
Questionnaires	Task analysis follow up studies	>30	Exposes subjective preferences	Dictates pilot work to avoid misunderstandings
Interviews	Task analysis	5	Flexible in-depth attitude & experience probing	Time consuming difficult analysis and comparison
Focus groups	Task analysis user involvement	6-9	Group dynamics	Difficult analysis low validity
Logging use	Follow-up studies	>20	Runs continuously exposing un/used features	Huge data masses need for analysis software & privacy infringements
User feedback	Follow-up studies	hundreds	Manages user requirements and view changes	Organization and manpower needed for analysis

Table 6.2 Test Methods [JN94]

6.3.2 Less is more – Nielsen's Graph

The main reason for developing the framework and its user interface can be summarized as the provision of a usable tool for supporting the development of structured bibliographic descriptions in and for managing digital archives. In other words, usability matters the most and to this end the usability evaluation of such a small special interest group is better governed by the discount usability engineering method and Nielsen's graph for usability testing, the "pay off ratio" [JN94] governing the number of test users necessary to obtain effective usability testing results.

This graph illustrated in the figure below, will serve as the guideline for evaluating the framework at hand and for the recruitment of sufficient test users for the user testing stage [JN93].

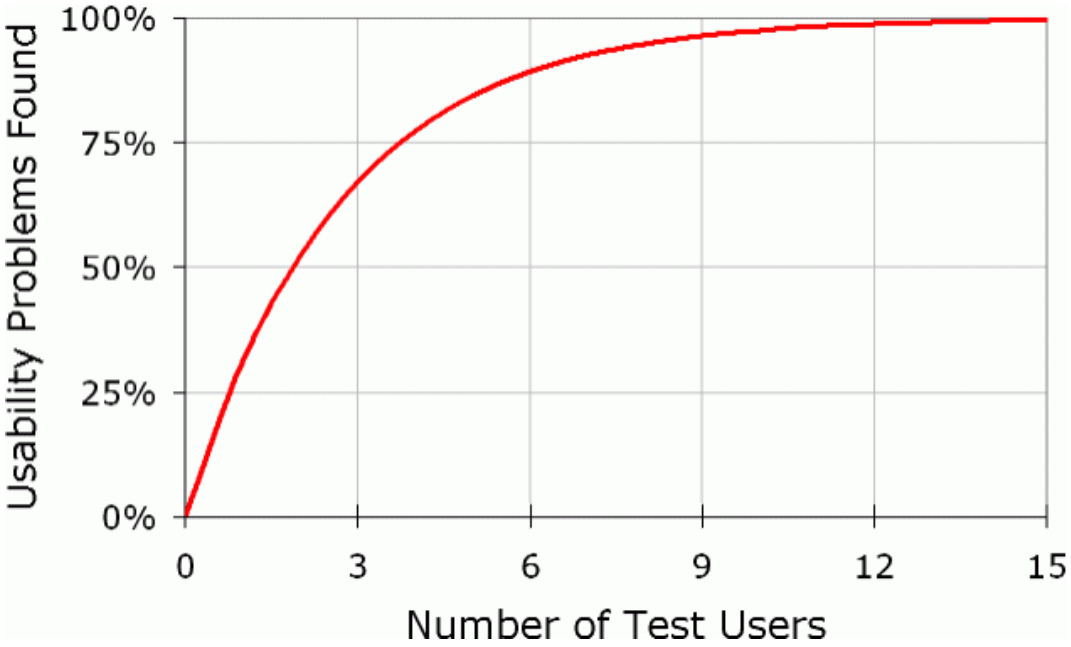


Fig. 6.2 Nielsen's Graph

A flashback to the introduction of the evaluation chapter outlines the different kinds of evaluation procedures, their types and what they aim to achieve. For the metadata creation framework developed in this dissertation, the graphical user interface, as the unit enabling access to the framework features is the subject of the evaluation. Hence the evaluation aims, test types and procedures are dictated by user interaction aspects represented by the user interface and putting the usability and user experience characteristics of the framework in the spotlight. Having listed the aims of evaluating the

user interface in addition to the description of the user testing and its associated characteristics in the sections 6.1 and 6.2 respectively, we now go on to have a look at the actual tests themselves and elaborate upon their planning and implementation in detail.

6.4 Test Goals and Plan

The metadata creation framework poses a solution to the digital archiving problems facing archivists managing scholarly archives in the humanities. As such the goals of the framework and indeed the graphical user interface are to ease interaction and abolish direct XML encoding replacing the relevant opening and closing tags by persistency tables represented by a relational database. The consequence of the proposed solution is then the usability and user experience evaluation testing how effective the user is assisted in the abstract metadata encoding process. It is these consequences that now dictate the test goals and test plans for the evaluation emanating from user tasks and encoding goals and summing up to a usability test of the graphical interface components. The components reflect the encapsulated encoding sections classified according to the metadata to be created. Reflecting upon the metadata heterogeneity and the archiving tasks dealt with in the chapters 2 and 3 respectively, the user interface component classes can be identified to be descriptive record collection interface, digital facsimile metadata interface and text summaries as created as encoded text.

As such the test goals of the summative evaluation will present the test user with the respective set of swing components for descriptive information defined by these categories and hence provide the media for providing the content of the opening and closing tags of the associated XML elements and hence the digital archiving entities. The ultimate goal is then to test the user acceptance of the graphical interface components as simple, comprehensive and learnable tools taking care of their digital archiving requirements. This ultimate goal can be outlined in detail in line with the discount usability engineering method and its set of usability heuristics in the introductory section 6.1 of chapter 6 prior to this section.

- Easy to learn
- Efficient to use
- Easy to remember

- Few errors
- Subjectively pleasing

6.4.1 Test Plan

Having set out the goals, an assessment of these goals has to be accompanied by an adequate plan for eliciting the empirical information. To this end the evaluation experiences of the formative evaluation and associated assessment of the test users comes in handy. The discount usability engineering method and Nielsen's Usability graph recommend a set of five heterogeneous test users, the heterogeneity being defined within the context of the three dimensions of user categories illustrated in the prior figures of this chapter 6. With scholarly archives representing a small special interest group within the field of humanities together with Nielsen's theory of a usability saturation point with increase in test users a set of five test user digital archivists seems reasonable for the summative evaluation. The set of five test users are all involved with archiving, digitization or edition activities within their archives and are associated with academic institutions:

- Alliance Israélite Universelle project associated with Steinheim-Institut
- Duisburger Institut für Sprach- und Sozialforschung
- Epigraphik at Steinheim-Institut
- Hegel Archiv
- Jonas Cohn Archiv

The test users represent an estimated age group between twenty and sixty-five of which the individual age of none of the test users was directly determined for the purposes of discretion. Although the test users and archivists in general tend to consider themselves as *"luddites"* observations during the formative evaluation revealed differences in computer experience and XML domain experience ranging from zero to tagging and XML vocabulary experts. On the other hand the archival domain experience showed extensive expertise in the humanities research content despite varied levels of experience illustrated by the qualification and academic titles summarized as follows:

- Graduate
- Research assistant

- Post doctorate
- Research Associate
- Associate Professor

6.5 Implementation

In addition to the formal classifications, the set of selected test users illustrated individual usage and test environment characteristics as predicted by the discount usability method. Whereas some users avoided dealing with questionnaires, others selected their own quantification and semiotics. Archivists working with text editions tended to add their own free text or express their preference to interviews as opposed to questionnaires. On the other hand archivists who also lecture opted to give scaled marks to each evaluation point. Taking all of this into consideration, the evaluation tests were planned to resemble a hybrid discount usability test combining user observation, focus group interviews to extrapolate and test the heuristics and the usability of the prototype metadata creation framework in general and its associated user interface components in particular. The resulting assessment then quantified the heuristics summarized in section 6.1 and above in accordance with the discount usability engineering method and Nielsen's graph and the implementation described in the succeeding section 6.4

Evaluation Criteria

In general the summative evaluation seeks to verify the metadata creation framework's usability within the digitization process. This usability is accompanied by and measured in accordance with the user acceptance and the capability to identify the individual archival needs within the interaction framework. However the evaluation criteria motivating the dissertation problem aim to motivate structured archiving and promote the framework and its abstraction as digitization infrastructure acceptable to the archivists. These evaluation criteria are therefore integrated into the discount usability engineering method resulting in an appropriate test design assessing the following evaluation criteria:

- Criteria 1: Structured archiving
    - Does the framework support structured digital archiving
    - Do the archivists accept the standardization constraints

- Criteria 2: User acceptance
    - Do the archivists recognise their work in the framework
    - Is the framework usable

The test design sees the test users accomplish standardized archiving tasks using their own artefacts and content on the basis of chosen metadata standards. Given that the structuring is guaranteed by the standard, the purpose of the test is to determine whether the archivists accept the structural dictatorship imposed on them by the metadata standard. The second criteria measured the framework usability on the basis of predetermined heuristics as outlined by the Nielsen's [JN93] discount usability engineering method. The empirics of the summative evaluation were then measured in two ways namely:

- Objectively: by assessing the heuristics associated with the framework use
- Subjectively: by eliciting user feedback on perceived usefulness and satisfaction

To this end the data collection took place at the test users' and their respective archives premises, however on an arbitrary workstation other than the users' common workstation. The theoretical background of the heuristic evaluation has been elaborated above in the preceding subsections outlining the measurable heuristics and the necessary test user groups. With these heuristics as the subject of the summative evaluation, measuring them and integrating them into the discount usability method is of importance for the evaluation implementation. Reflecting upon the formative evaluation and the characterization of the test users described above influences the test design and the data collection procedures. Since the test users had the tendency to avoid system functional questions preferring aesthetic and perception the heuristics evaluation data is by observation categorized according to the criteria outlined in the test goals.

Evaluation Results

Given the test user categories and their characteristics summarized above the logical consequence is a set of flexible user interviews based on typical user tasks tested using the developed framework. The test users were subject to a documented observation and a heuristic questionnaire on the subjective usability of the metadata creation framework.

The main reason for combining the two collection procedures being user tendency to restructure unfavourable questions or test procedures as experienced in the formative evaluation. Furthermore, with the observation protocols and procedures deviations in the evaluation metrics are simplified and mostly confined to the questionnaire and its subjective analysis. With the set of test users respecting Nielsen's graph their characteristics were bound to illustrate differences in answering the evaluation question. To this end archivists with teaching activities tended to give grades to each section whereas those mostly into editions preferred to write commentary answers. The results of the survey are illustrated in summary below and give an insight to the proof of concept, acceptability and the usability of the developed metadata creation framework.

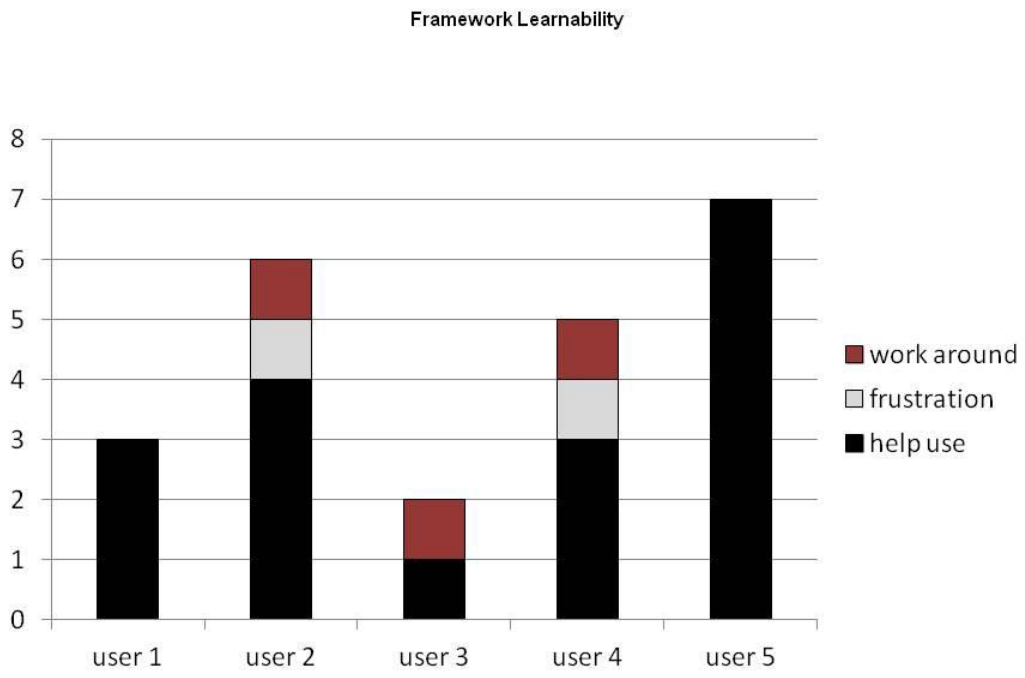


Fig. 6.3 Learnability Heuristics Graph

Fig. 6.3 above summarizes the learnability heuristics results of the evaluation. Whilst the performance measures are outlined below the detailed description and interpretation follows on page 210.

Performance Measures:

- Work around: number of times user achieved tasks using alternative routes

- Frustration: number of times user expressed frustration
- Help use: number of calls for assistance during a task

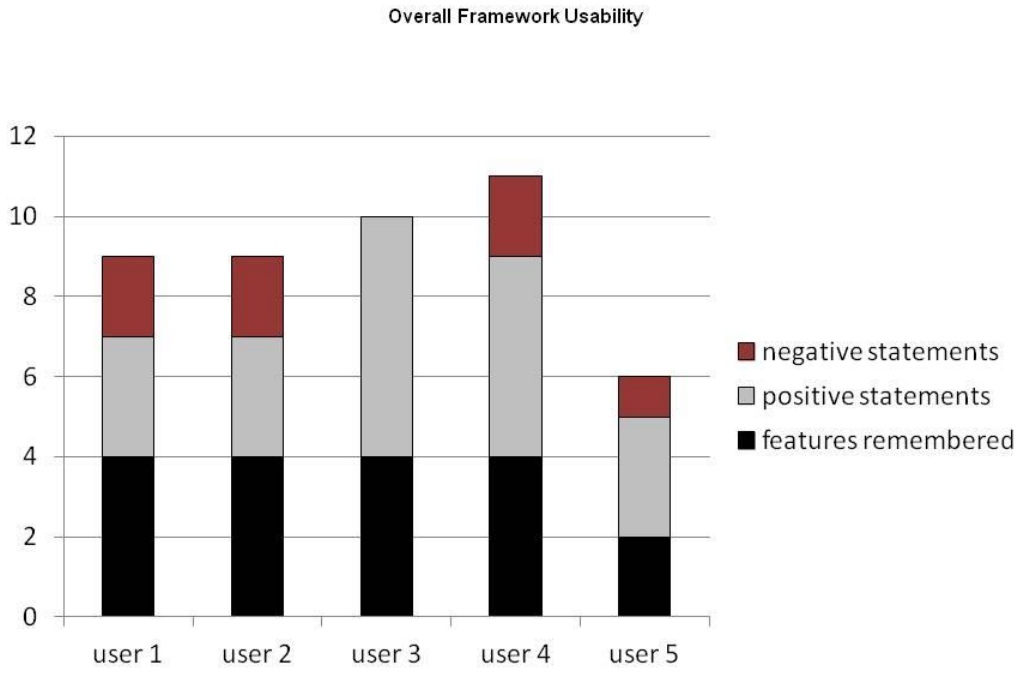
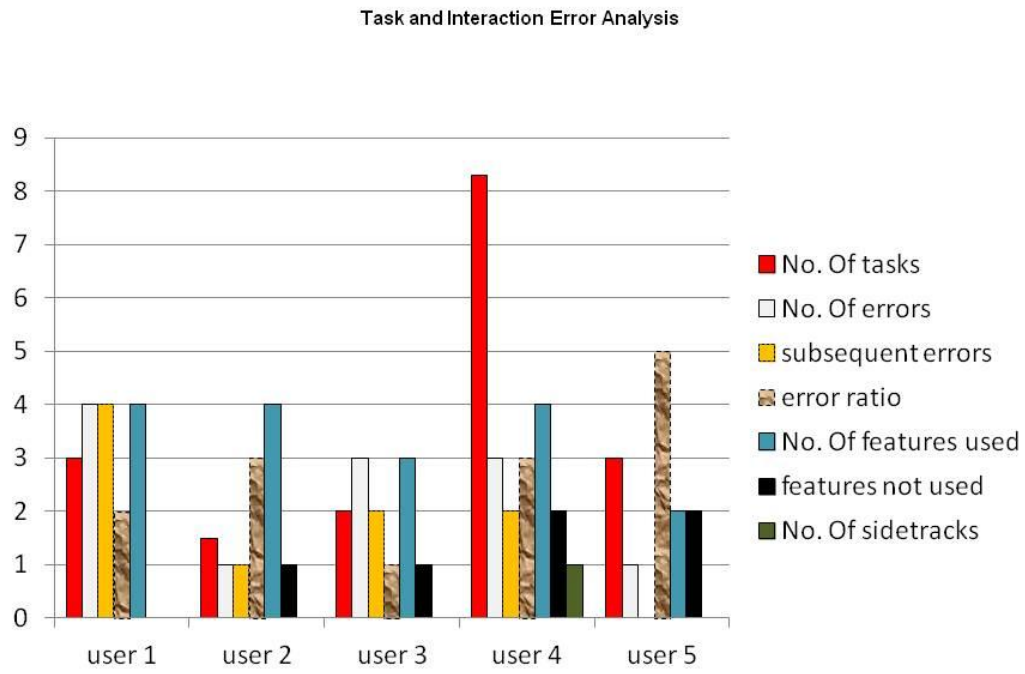


Fig. 6.4 Usability Feedback

Fig. 6.4 above summarizes the usability feedback results of the evaluation on the basis of the subjective measures outlined below. The detailed description and interpretation follows together with that for Fig. 6.3 on page 211.

Subjective Measures:

- Negative statements: number of negative statements
- Positive statements: number of positive statements
- Features remembered: number of features remembered



6.5 Framework and Interface Error Rate

Fig. 6.5 above summarizes the results of the framework and interface error rate tests. The respective performance measures are outlined below complementary to the description and interpretation on page 211.

Performance Measures:

- No. of tasks: number of tasks achieved within the specified test time
- No. of errors: numbers of errors made including incorrect choices in dialogue boxes and wrong menu choices
- Error ratio: ratio of errors made in relation to the number of tasks achieved within the specified test time
- No. of features used: number of features actually used in the business case
- Features not used: number of offered features not used in the business case
- No. of side-tracks: Enumeration of distractions and activities outside the business case

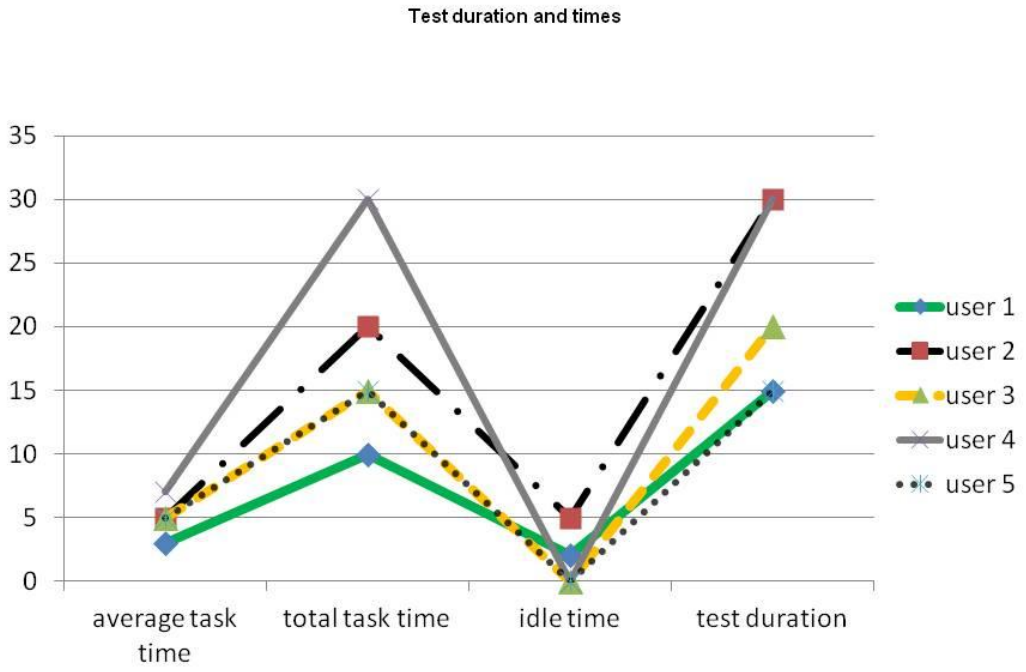


Fig. 6.6 Time Based Framework Usability Results

The time usage metrics of the framework were measured according to performance measures outlined below and the results of which resemble Fig. 6.6 and Fig. 6.7. During the test the interface and hence the framework were assessed with respect to their contribution towards efficiency during the digitization process.

Performance measures:

- Average task time: average time user required to carry out a task in general
- Total task time: total time required by each user to carry out the same task
- Idle time: task time spent without active user interaction with system
- Test duration: amount of time required by the individual user to complete test tasks

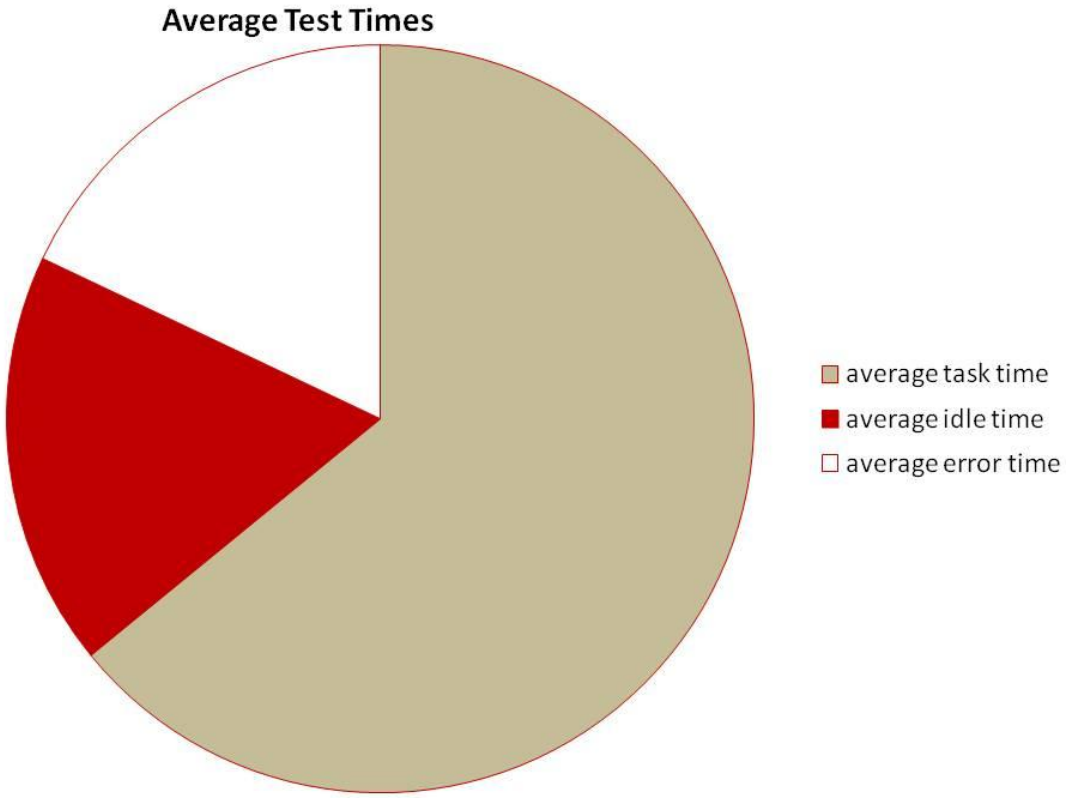


Fig. 6.7 Overall User pre-occupation with Framework and Task

Summary of Performance Measures

- Average task time: average time spent on task in general
- Average idle time: average task time where system is idle
- Average error time: average task time spent on errors
- A descriptive of the relevant empirical subjective measures includes:

The usability testing measures in Fig 6.8 and Fig 6.9 are qualitative subjective measures based on the subjective judgment of the individual users. As such the usability test results reflect report on the frequencies i.e. how many test users reflect a particular judgement towards a feature or the system in general. The measures in Fig 6.8 refer to the consistency, acceptance of the system in general as well as menus and labels in particular. The subjective measures in Fig 6.9 on the other hand describe:

- Learnability: ease of learning the system

- Effort: entry levels of effort required to use the system
- Simplicity: ease of doing a particular task with the system
- Confusing: ease of understanding navigation and interactions
- Help: usefulness of helping aids and navigators
- Completeness: task coverage and applicability beyond test scenario
- Comprehensive: purpose coverage

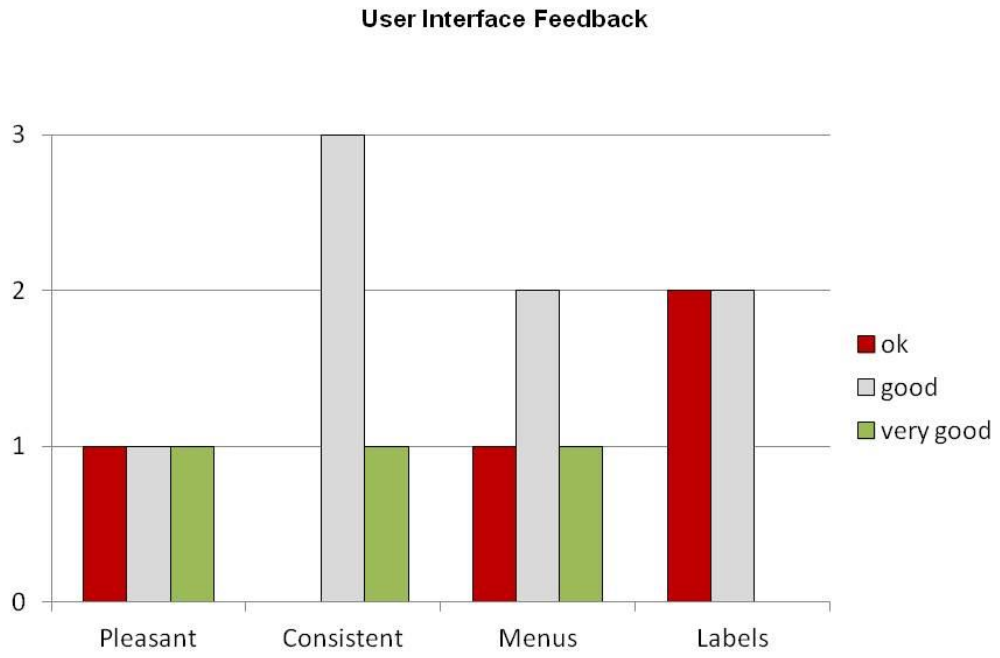


Fig. 6.8 Test User Feedback on Interface

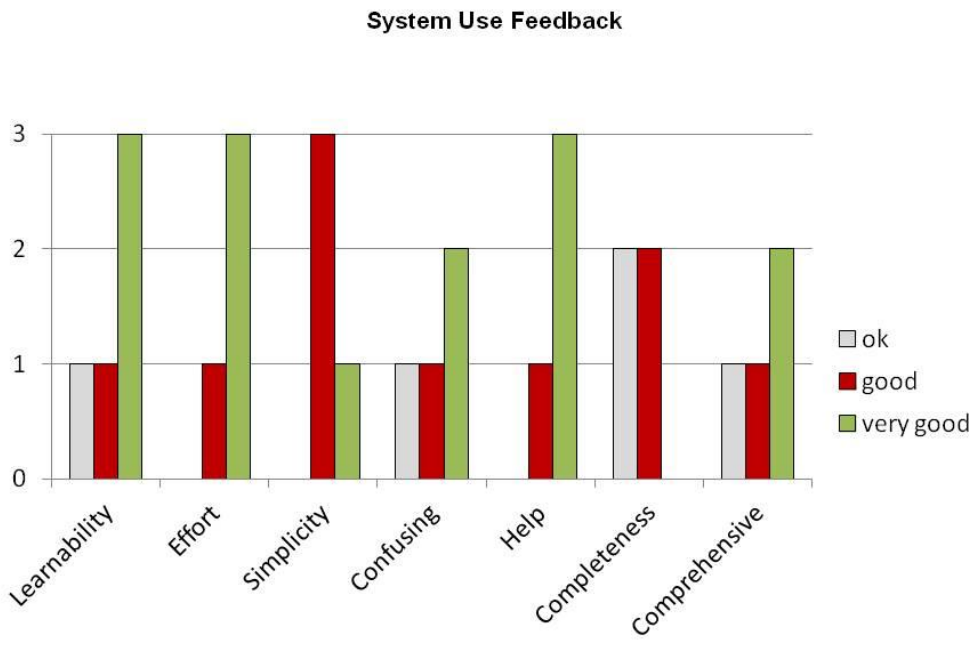


Fig. 6.9 User Feedback on Heuristics

6.5.1 Interpretation of Results

The evaluation results illustrated in the preceding subsection outline and assess the effectiveness and the response of the target user group to the proposed metadata creation framework. An analysis of this effectiveness and the information content of the results is a key aspect of the dissertation tackled here in the form of an interpretation as presented.

Learnability

The learnability heuristics graph focussed on the following three characteristic heuristics to assess the target users' capability to intuitively learn how to use the graphical user interface framework.

- Work around
- Help use
- frustration

The evaluation results illustrated in Fig 6.3 attest the metadata creation framework and its associated graphical user interface positive learnability. The majority of the test users had either no difficulties using the framework system and when difficulties were met, the offered helping mechanisms were well utilized after exhausting personal workarounds. The general implication being that most users showed familiarity with the implemented technology and had enough self-confidence and trust to adapt individual aspects and solve trivial problem encounters. Nevertheless, a small section made very little use of workarounds and offered help however not showing any signs of frustration or a denial of acceptance towards the framework system.

Usability Feedback

The graphical user framework usability and aspired swing interface awareness evaluated along the criteria illustrated in Fig. 6.4 looked at user statements in relation to the interface framework. The statements were categorized as:

- Negative Statements
- Positive Statements
- Features Remembered

The user acceptance and interface awareness aspects influenced by these characteristics relay an overall acceptance of the metadata creation framework. The ratio of features remembered overshadows the other criteria and are in line with the expected result particularly with respect to the user interface. The positive and negative statements seem at par, reflecting optimization potential whilst acknowledging the high acceptance level.

Task and Interaction Results

In addition to the general heuristics concerned with the interface, awareness and user recognition of implemented features, the results of the task related navigation and efficiency aspects of the framework and its interface analysed on the basis of the criteria outlined below, are illustrated in Fig. 6.5

- Number of tasks
- Number of errors

- Subsequent errors
- Error ratio
- Number of features used
- Number of features not used
- Number of side tracks

The overall number of tasks carried out over a specified time slot varied in relation to required features notably user 2 and user 4. Nevertheless, the number of used features overwhelmed the features not used in addition to the relatively low error ratio. All in all, the results reflect the interface framework's contribution towards efficiency and error reduction in the metadata creation process. However, individual peak values for error rate and number of tasks used indicate that this the effectiveness of the framework interface is not evenly distributed across the board. The time analysis on the other hand conclusively reflects low idle times and concentrated focus on the part of the framework users as well as low average times necessary to complete given tasks. In other words, the framework improves and increases user concentration on the task and implicitly an awareness of the metadata structuring process. The error time in Fig. 6.7 illustrates the average time in which the test user was preoccupied with errors.

User Experience Feedback

The remaining graphs in Fig. 6.8 and 6.9 outline the subjective user response and attitude towards the metadata creation framework interface and implicitly the structured metadata creation process and associated data interchange in general. The evaluation heuristics and the results witness a very high level of acceptance in both usability and comfort as well as in the possible implementation of the framework as a digital archiving instrument thereby confirming the concept proposed in the dissertation.

6.5.2 Summary

A summative evaluation of the metadata as implemented above symbolized the final stage of the dissertation giving an insight to the necessity and the effect of the framework for and on the target user groups. An assessment of the user interface features with respect to the usability and in light of the heterogeneous XML encoding is represented by

the summative evaluation whose implementation and structure is based on the evaluation theories outlined in the earlier sections of chapter 6. This evaluation summarized the results of the usability tests validating them against proposed hypothesis of simplifying the process of creating structured metadata for integrated digital archiving as outlined in chapter 1. All in all the overall framework usability shows a positive user resonance supported by the high ratio of positive statements in comparison to negative statements expressed by the test users.

This outcome is supported by the test aspect in the form of the features remembered by the test users which is almost constant hence approving of the resonance determined above. However the use of help mechanisms and the manual were also rather frequent despite the low rate of frustration and workarounds. This could be attributed to the novelty of the tool despite common user interface and menu graphics, especially due to the fact that some of the test users were not structuring their archives at all. The effectiveness of the metadata framework and the minimum error rate also pointed towards a favourable usability as is summarized by the average time the user allocated to the metadata creation tasks. In addition to the usability and suitability as a markup tool, the framework was also criticized by users most probably involved in manual mark up. The general argument against automated markup was the reference to digital editing and textual encoding of non-standardized text and document structures constantly requiring tailor made solutions. However, the emphasis of this dissertation has indeed been on standardized markup and archival structure as a step towards machine readability and as infrastructure for archival interoperability. The latter require and are governed by standardized recommendations such as the resource description framework for the semantic web or the open archival information system OAIS.

All in all the formative and summative evaluations prove the notion of abstract XML metadata creation outside the manual XML encoding sphere and in a standard usable day to day graphical user interface not requiring XML encoding knowledge. The fact that experienced encoders and inexperienced encoders alike welcomed the framework as a plausible archiving infrastructure supported the goal of encouraging archivists to structure their archival data.

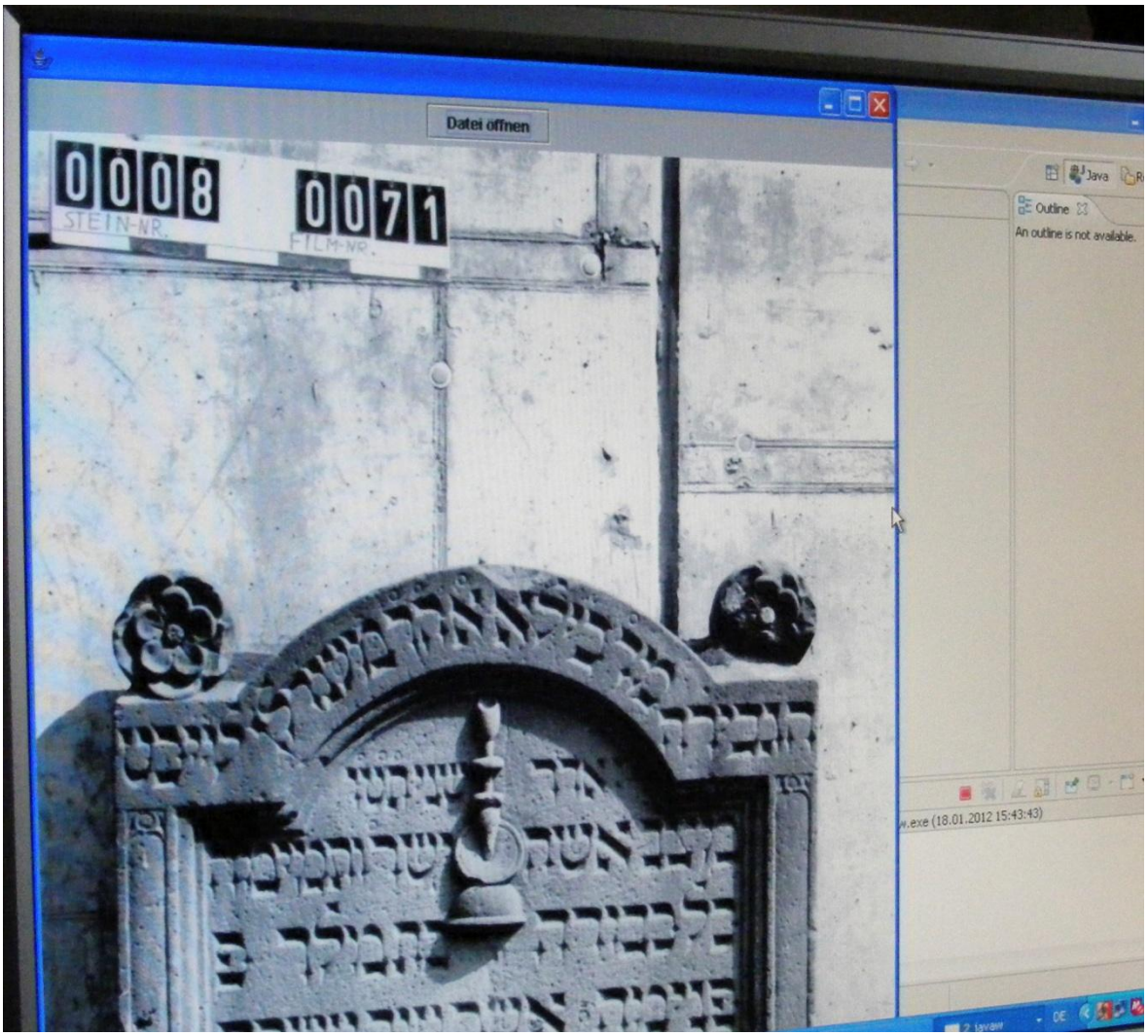


Fig. 6.10 Test Image upload during evaluation

Fig. 6.10 illustrates a usability problem discovered during the tests using the discount usability method. In this case digital images of archived material were uploaded and preserved in very high resolution to counter problems of illegibility. As a result uploaded images needed to be scaled down to match required performance during access whilst adapting the framework and it's upload facilities to the digital input.

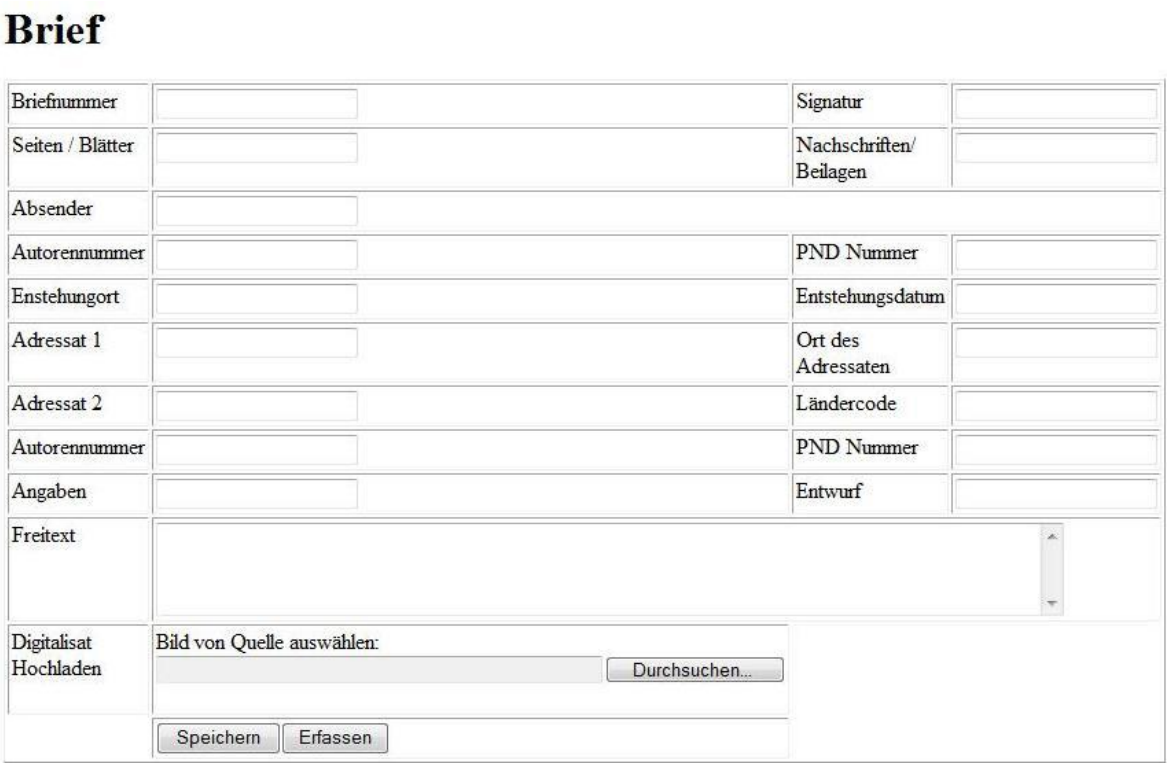


Fig. 6.11 Interface for correspondence artefact metadata entry

Fig. 6.11 above shows an illustration of the test user interface for collecting correspondence artefact metadata. The interface also provides for the associated uploading of the corresponding digital image object. The combination of structured data collection and the direct association of a digital resource proved crucial for the user awareness of the relationship between metadata created and the object viewed in association.

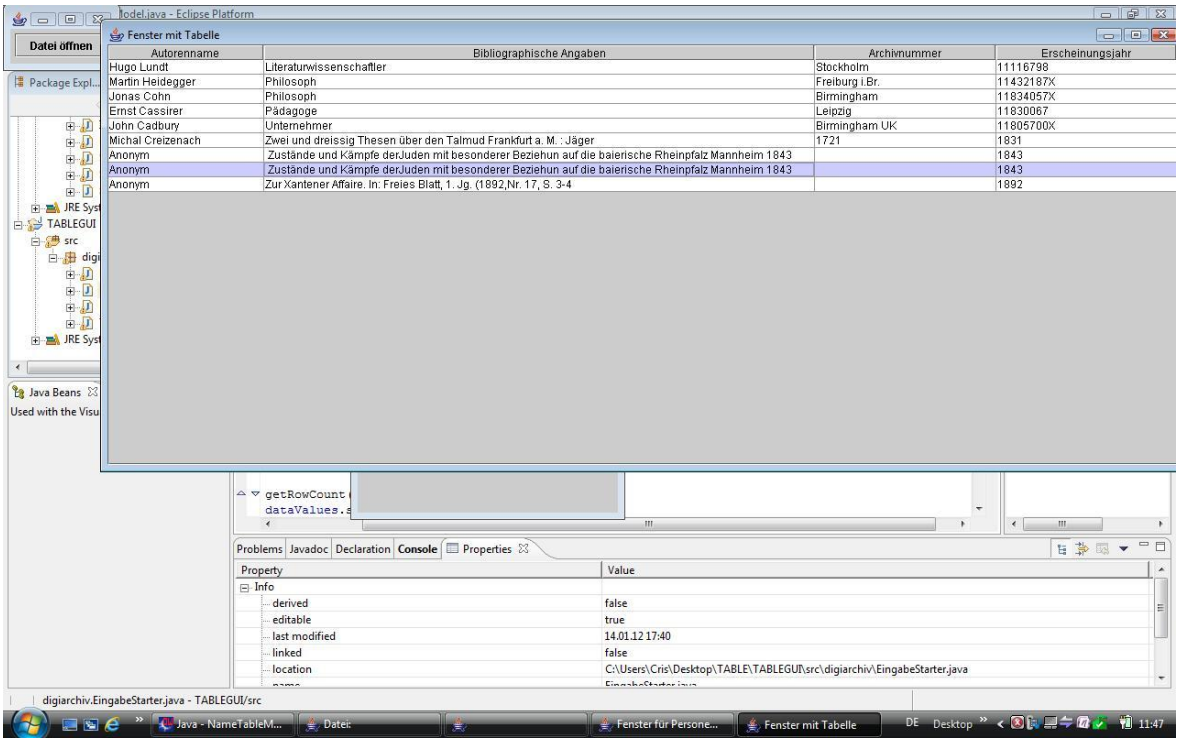


Fig. 6.12 Test interface for author metadata collection

The Fig. 6.12 above serves to illustrate the test environment during the development and assessment of the swing user interfaces for capturing bibliographic author metadata. The testing within the framework of the development environment enabled a rapid correction of discovered interface problems and errors. In the example above the test user oversaw data capturing titles and proceeded to capture author data in the user's standard sequence. As a result the data capturing sequence common to the test users influenced the sequence of the data capturing interfaces of the framework.

In conclusion, the illustrations above serve to visualise the user test scenarios with respect to the evaluation and in relation to the evaluation aims. The evaluation notions implemented focussed in addition to assessing the acceptance of the framework as part of the structured digitization process. On the other hand problem discovery coupled with the interface development served the implementation of a usable and acceptable technological intermediary.

7 Conclusion and Outlook

In this dissertation I have presented my work on a conceptual graphical user interface supported framework aiding the creation of heterogeneous XML metadata for digital archives based on metadata requirements of the digital archivists and their integration within the digital archiving rules (RNA) [WK10] and the functional requirements for library and archive information systems [FR08]. The proposed framework approach is meant to encourage archivists especially those from the humanities to structure their digital archives with the help of a usable graphical user interface framework. Due to the digital nature of the archives, the required metadata heterogeneous or otherwise are to be structured using XML and the structuring is then implemented within the framework with the help of state of the art XML data binding techniques and java with its related swing components for interface design and implementation. The prototype has been realized using the java swing components and an XML binding schema which matches metadata stored in a persistence represented by a relational JBOSS database to selected XML standardized schema specified by XML Schema Document (XSD) files. The implementation of the proposed problem solution and the prototype were accompanied by a formative and a summative evaluation in the proposed users' working environments eliciting standard digital archivists' tasks and accessing the adequacy of the framework as a usable tool for solving the XML metadata creation problem. In addition to confirming the adequacy and usability of the framework, the evaluations and associated user tests revealed standard user types for digital archiving in the humanities hence defining the necessity and extend of the testing process with respect to user experience and functionality and the relevance of the resultant metadata to the digital archiving process and its interacting actors.

7.1 Summary

The work presented in the preceding chapters illustrated the challenges associated with digital archiving outlining the record collecting aspects, the impact of digital facsimile and the resultant heterogeneity of the metadata and the semantic considerations as the digital archives are presented as online documents subject to technological and behavioural dictates of the internet domain. With the implementation of XML data binding with java, the dissertation brings in a popular, platform independent and widely implementable

technological aspect into the digital archiving sphere. Java in its enterprise edition form provides a wide range of programme classes and libraries easing the software development process and allowing software and pattern reuse within the software development process. Together with java, XML metadata can now be abstractly modelled and represented by structural patterns independent of semantics and mark-up syntax relieving the digital archivist from the field of humanities from non-core, time consuming and training intensive activities. The structuring of the metadata and the record collection activities can now be channelled via graphical interfaces as swing components, abstract window toolkit components or otherwise ensuring reuse and technological encapsulation separating the user interface from the data being processed, a principle notion of XML.

In addition to the technical aspects and the specified digital archiving problem cited as the research question, the framework and the XML binding techniques dealt with and documented in the dissertation also bridge the gap between digital archiving and computing in the humanities. With the general end product of a digital preservation or archiving task being a web document for online presentation and research, the isolation of digitization in the humanities from mainstream computing is overcome with the implementation of java and its model view controller architecture in addition to the n-tier middleware to complement the resultant DOM modeled presentation documents. The structuring and preservation mark-up which is of importance to the digital archivists is preserved despite being encapsulated within the framework defined by a swing based user interface mask or any other interaction module embedding within the java architecture and technology.

Cooperation between mainstream computing and information sciences and the humanities is key to the future of both areas of research as both are dependent upon each other. Whilst the former obviously provides for the infrastructure for state-of-the-art access to and dissemination of information, the latter provides the content which utilizes the infrastructure.

In addition to providing the content, access to archival text and object artefacts enables broader research in the humanities field in question whilst reaching a wider a more diverse audience. The result of such a research audience is wider application areas of the research material and broader perspectives as opposed to local common perspectives.

The notion of a broader audience is noted in the prototype scholarly archive represented by the Jonas Cohn Archive where correspondence and material stretch from the east to the west of continental Europe and the United Kingdom and from Asia across the Orient right up to the Americas a mirror of the Jewish Diaspora reflected by the archive content.

Digital Humanities as the novice term to this interdisciplinary field involving the mark-up and structuring classical works of the humanities is tasked with a range of challenges some of which have been dealt with in this research work. These challenges are derived from the human machine interaction nature of the systems required to implement such digital humanities activities. As such considerations on technical standardization and structural consolidation coupled with user experience aspects are key to tackling challenges faced by digital humanities in a rapidly changing technological environment. On the other hand user evolution and future relationships between researchers in the humanities and technological devices must not be underestimated. Not all researchers in this field are *"luddites"* and therefore future system interactions should take this into consideration offering appropriate interfaces and space for further user participation in the archiving process. The user interfaces implemented in the metadata creation framework introduced here focused on the state-of-the-art java swing interfaces obviously subconsciously common to modern day computer users irrespective of their background. The results of the summative evaluation and the fact that the users recognised menus and graphical surfaces without extensive assistance support this fact.

To sum up the JAXB supported framework approach managed to address the dissertation question providing for the graphical user interface through java and swing whilst enabling an encoding knowledge free metadata encoding. The semantic heterogeneity irritable even to manual encoders was replaced by the centralized record element collection entity and the multiple schemas realized by marshalling java content objects to the XML structure. The bibliographic and semantic web considerations involved in digital archiving can now be dealt with in an elegant way open to new structures and schema in the future. The key is the binding schema and the possibility of creating and representing object content structure within a metadata framework. With this infrastructure in place archivists can tackle any XML structured Schema and crosswalk their existing records.

Contribution

- Framework provides a basis for information interchange and interoperability amongst digital archives
- Contributes towards structured encoding of digital archives and open access to digital archives
- Consolidation of XML digital archive metadata outside the scope of a further XML standard avoiding multiple encoding and code-lists
- Supports digital archiving activities without XML encoding knowledge
- Resembles an acceptable mediator between archivist requirements and technical requirements for preservation and interoperability

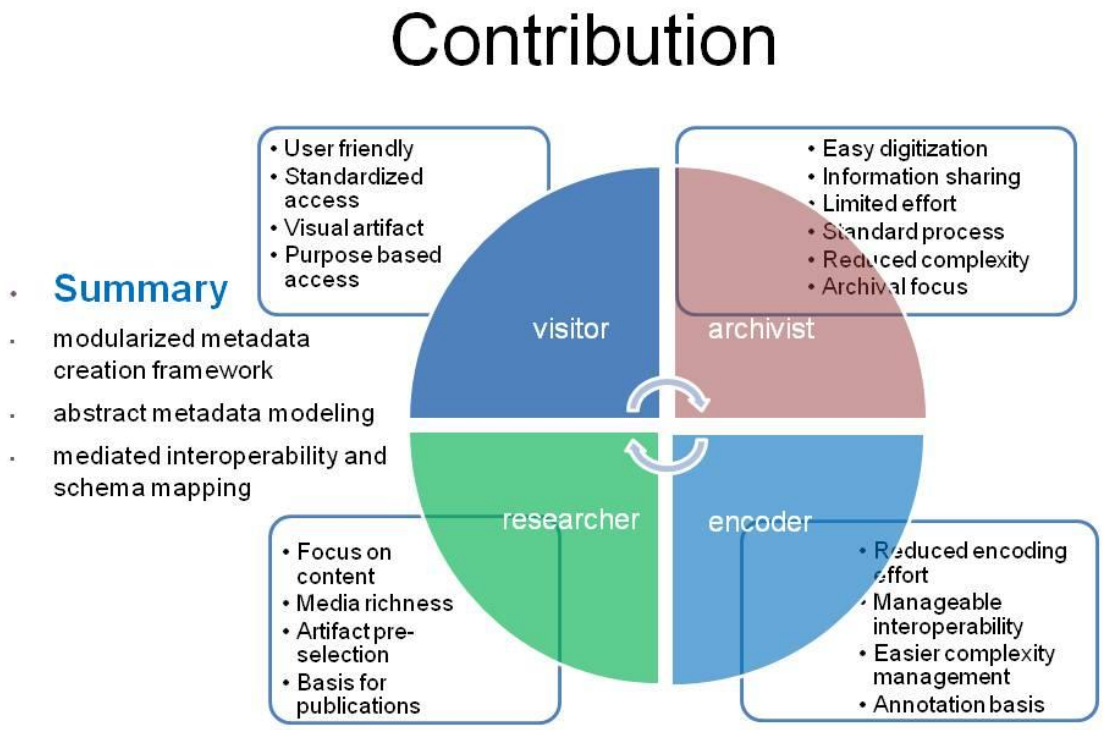


Fig. 7.1 Dissertation Contribution

7.2 Outlook

This notion of XML data binding frameworks for digital archives may be extended to include broader metadata descriptions such as those involved in digital editions. Although digital editions involve the definition of a wide range of description elements as already illustrated by TEI, it remains a matter of voluminous processing most probably easier to process with future computing machines as they also rapidly improve computing speed and efficiency. Furthermore, such a binding schema oriented abstractions may also be extended to developments in the semantic web sphere provided an environment with standardized description elements. A further improvement of the metadata creation framework could be an extension to other programming languages providing for graphical user interface development and XML data binding alike. This would minimize the dependency on the java programming language whilst maintaining the standard interfaces for optimal usability. A further application area could be the digital museums which similar to digital archiving is closely associated with bibliographic descriptions and person entity relationships. Generally the framework can be useful for any scenario involving standardized metadata creation for any non-programmer or XML expert user community.

Bibliography

[AC80] Universitätsbibliothek Braunschweig. *allegro-C Bibliothekssoftware – konfigurierbares Datenbanksystem*
http://www.allegro-c.de/
accessed 10.05.2012

[AD07] Di Iorio, A. (2007). *Pattern-based Segmentation of Digital Documents: Model and Implementation*
http://www.cs.unibo.it/pub/TR/UBLCS/2007/2007-05.pdf
accessed 20.08.2011

[AG92] Andleigh, P., Gretzinger, M. (1992). *Distributed Object –Oriented Data-Systems Design.* Prentice Hall

[AH97] Universitätsbibliothek Hamburg. *allegro-HANS Bibliothekssystem für Handschriften, Autographen, Nachlässe und Sonderbestände*
http://www.sub.uni-hamburg.de/bibliotheken/fuer-die-fachwelt/allegro-hans.html

[AT06] Archivists'Toolkit Archival data management system
http://archiviststoolkit.org/node/96
accessed 01.08.2011

[B05] Bertot, J.C. *Assessing Digital Library Services: Approaches, Issues, and Considerations*
http://www.kc.tsukuba.ac.jp/dlkc/e-proceedings/papers/dlkc04pp72.pdf
accessed 24.07.2011

[BA05] Blandford, A, Adams, A, et al. (2008).*The PRET A Rapporter Framework: evaluating Digital Libraries from the perspective of information work*. Journal Information Processing and Management: an International Journal archive Volume 44 Issue 1, January, 2008 Pages 4-21

[BB02] Consultative Committee for Space Data Systems. (2002) .*Reference Model for an Open Archival Information System (OAIS)*

[BB07] Bomsdorf, B. (2007).*The Web Task Model Approach to Web Process Modelling.* Springer-Verlag

[BD07] Daum, B. (2007). *Java 6 Programmieren mit der Java Standard Edition*. Addison Wesley

[BE06] Bruijn, J. Ehrig, M. et al. (2006) *Ontology meidation, merging and aligning*

http://disi.unitn.it/~p2p/RelatedWork/Matching/mediation-chapter.pdf

accessed 02.05.2011

[BLVU03] Bollen, J., Luce, R. et al. (2003). *Usage Analysis for the identification of Research trends in Digital Libraries*. D-Lib Magazine May 2003 Volume 9 Number 5

http://www.dlib.org/dlib/may03/bollen/05bollen.html

accessed 10.05.2011

[BO06] Botterweck, G. (2006).*A Model-Driven Approach to the Engineering of Multiple User Interfaces*

http://ulir.ul.ie/bitstream/handle/10344/2145/2006_Botterweck.pdf?sequence=2

accessed 10.07.2011

[BM02] McLaughlin, B. (2002). *Java and XML Data Binding*. O'Reilly Media

[BS07] Buschmann, F., Schmidt, D. et al..(2007). *Pattern-Oriented Software Architecture – On Patterns and Pattern Languages.* John Wiley & Sons

[BV03] Bishop, A. P. et al.. (2003). *Digital Library Use – Social Practice in Design and Evaluation*. MIT Press

[CH02] Choudry, S., Hobbs, B. et al.. (2002). *A Framework for Evaluating Digital Library Services*. D-Lib Magazine July/August 2002 Volume 8 Number 7/8

http://www.dlib.org/dlib/july02/choudhury/07choudhury.html

accessed 02.05.2011

[CR06] Cohen, D. Rosenzweig, R. (2006). *Digital History A Guide to Gathering, Preserving, and Presenting the Past on the Web*. University of Pennsylvania Press

[DA10] Das Bundesarchiv: Digitalisiertes Archivgut in Online Findbüchern http://www.daofind.de/

accessed 03.07.2011

[DC09] Dublin Core Initiative

http://dublincore.org/

accessed 10.07.2011

[DE10] Dappert, A., Enders, M. (2010).*Digital Preservation Metadata Standards*
http://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf
accessed 10.03.2011

[DF10] DFG Praxisregeln “Digitalisierung”
http://www.dfg.de/download/pdf/foerderung/programme/lis/praxisregeln_digitalisierung.pdf
accessed 10.02.2012

[DR99] Dumas, S., Redish, J. (1999).*A Practical Guide to Usability Testing*. Intellect Ltd

[DV10] DFG Viewer: Profil der Metadaten
http://dfg-viewer.de/
accessed 04.03.2011

[EAD09] Encoded Archival Description
http://www.loc.gov/ead/
accessed 09.11.2011

[EG95] Gamma, E., Helm, R. et al..(1995). *Design Patterns – Elements of Reusable Object-Oriented Software.* Addison-Wesley

[EM79] McCormick, E.J., Sanders, M. (1993). *Human Factors Engineering and Design*. McGraw Hill

[F07] Fuhr, N. et al. (2007). *Evaluation of Digital Libraries*. Springer-Verlag 2007

[FR08] Functional Requirements for Bibliographic Records
http://archive.ifla.org/VII/s13/frbr/frbr.pdf
accessed 05.03.2012

[FS09] Schwarz, F. (2009). Entwicklung, Anwendung und Evaluation einer internetbasierten Lernumgebung
http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-21875/schwarz_diss.pdf
accessed 09.05.2011

[FT07] Fuhr, N., Tsakonas, G. et al. (2007). *Evaluation of digital libraries*. International Journal on Digital Libraries archive Volume 8 Issue 1, October 2007 Pages 21 - 38 Springer-Verlag

[GH00] Hodge, M., (2000). *Best Practices for Digital Archiving – An Information Life Cycle Approach*. D-Lib Magazine January 2000 Volume 6 Number 1
http://www.dlib.org/dlib/january00/01hodge.html
accessed 20.02.2012

[GL00] Greenstone Digital Library Software
http://www.greenstone.org/
accessed 12.02.2012

[G08] Gilliland, A.J. (2008).*Setting the Stage.* Introduction to Metadata 3.0. J. Paul Getty Trust
http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.pdf
accessed 20.04.2012

[HL00] Hill, L, Carver, L, Larsgaard, M. et al. (2000). *Alexandria Digital Library: User Evaluation Studies and System Design*. Journal of the American Society for Information Sciences. 51(3): 246 – 259. 2000
http://www.icess.ucsb.edu/~frew/cv/pubs/2000_user_evaluation.pdf
accessed 10. April 2012

[JN93] Nielsen, J. (1993).*Usability Engineering*. Academic Press

[JN94] Nielsen, J. (1994). The Discount Usability Engineering Approach - Heuristic Evaluation
http://www.useit.com/papers/guerrilla_hci.html
accessed 10.02.2012

[KA03] Aasbo, K. (2003).*Dublin Core a useful tool for photography?*
http://www.agencia.cnptia.embrapa.br/Repositorio/aasbo_000finhqfiu02wyiv80z4s473s9i2309.pdf
accessed 20.05.2011

[KFT05] Klas, C., Fuhr, N. et al. (2006). *An Experimental Framework for Comparative Digital Library Evaluation: The Logging Scheme*. JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries
http://www.is.inf.uni-due.de/bib/docs/Klas_etal_06.html

accessed 10.02.2012

[L00] Langer, A.M. (2008).*Analysis and Design of Information Systems* Springer-Verlag

[LMB02] Lockee, B, Moore, M., Burton, J. (2002). *Measuring Success: Evaluation Strategies for Distance Education*. Educause Quarterly Number 1 2002

[MA04] MARC 21 XML Schema (MARCXML)
http://www.loc.gov/standards/marcxml/marcxml-design.html
accessed 08.01.2012

[ME10] Metadata Encoding and Transmission Standard
http://www.loc.gov/standards/mets/
accessed 08.01.2012

[MR09] Macefield, R. (2009). *How To Specify the Participant Group Size for Usability Studies: a Practitioner's Guide*. Journal of Usability Studies November 2009 Volume 5 Issue 1

http://www.upassoc.org/upa_publications/jus/2009november/JUS_Macefield_Nov2009.pdf

accessed 20.05.2013

[NH09] Neuroth, H., Oswald, A. et al.. (2009). *nestor Handbuch – Eine kleine Enzyklopädie der digitalen Langzeitarchivierung Version 2.0 Kapitel 4.2 – Das Referenzmodell OAIS*. Verlag Werner Hülsbusch

[NM01] Naiburg, E.J., Maksimchuk, R.A. (2001).*UML for Database Design*. Addison Wesley

[OM03] Ort, E., Mehta, B. (2003). *Java Architecture for XML Binding (JAXB)*
http://www.oracle.com/technetwork/articles/javase/index-140168.html#introjb
accessed 22.01.2012

[OR01] OCLC/RLG Working Group on Preservation Metadata (2001). *Preservation Metadata for Digital Objects*
https://www.oclc.org/resources/research/activities/pmwg/presmeta_wp.pdf
accessed 02.03.2012

[P99] Paternò, F. (2000). *Model-Based Design and Evaluation of Interactive Applications*. Springer-Verlag

[RA05] Assisi, R. (2005).*J2EE mit Eclipse 3 und JBoss – Enterprise-Anwendungen mit der Open-Source-Plattform entwickeln*. Hanser

[RF07] Radeke, F., Forbrig, P. (2007) *Patterns in Task-Based Modeling of User Interfaces,* Springer-Verlag

[RR05] Rosenthal, D., Robertson, T. et al. (2005). *Requirements for Digital Preservation Systems – A bottom up approach.* D-Lib Magazine November 2005 Volume 11 Number 11
http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html
accessed 10.04.2011

[QA10] Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/pmh/
accessed 20.05.2012

[S00] Svenonius, E. (2000).*The intellectual foundation of information organization*. MIT Press

[SJ10] Sauro, J. (2010). *A Brief History Of The Magic Number 5 In Usability Testing*
http://www.measuringusability.com/blog/five-history.php
accessed 20.05.2013

[SD79] Dworatschek, S. (1979). *Management Informationssysteme*. Walter De Gruyter & Co.

[TEI09] Text Encoding Initiative
http://www.tei-c.org/index.xml
accessed 20.05.2012

[TH02] Hillesund, T. (2002). *Many Outputs - Many Inputs: XML for Publishers and E-book Designers*
http://journals.tdl.org/jodi/article/viewArticle/76/75
accessed 09.03.2011

[TS04] Saracevic, T. (2004).*Evaluation of digital libraries: An overview*

http://comminfo.rutgers.edu/~tefko/DL_evaluation_Delos.pdf
accessed 11.06.2011

[TW85] Evaluation Strategies for library/Information systems
http://informationr.net/tdw/publ/papers/evaluation85.html

[UW05] Umstätter, W., Wagner-Döbler, R. (2005).*Einführung in die Katalogkunde Vom Zettelkatalog zur Suchmaschine*. Hiersemann-Verlag

[W3C] World Wide Web Consortium
http://www.w3c.org
accessed 10.03.2012

[WB03] Witten, I., Bainbridge, D. (2003). *How to Build a Digital Library*. Morgan Kaufmann

[WK10] Weber, J., Kaukoreit, V. (2010).*Regeln zur Erschießung von Nächlässen und Archive*
http://kalliope.staatsbibliothek-berlin.de/verbund/rna_berlin_wien_mastercopy_08_02_2010.pdf
accessed 10.03.2012

[WP99] Pardi, W. J. (1999).*XML in Action Web Technology*. Microsoft Press

[WR06] Wright, R. (2006). *What Archives Want – The requirements for digital technology*
http://tech.ebu.ch/docs/techreview/trev_308-archives.pdf
accessed 10.11.2011