Digital Archives as Part of Digital Humanities Research Infrastructure

Towards a Standardized Model of Archiving and Dissemination

Stigler, Hubert Johannes

Abstract - ***Due to the increasing degree of digitization in research, the topic of modeling scholarly content is gaining increasing prominence in Humanities and Cultural Studies. XML-based data formats appear well-suited to flexible, metadata-enriched forms of storage of textual data: the primary content of documents is augmented with additional descriptive elements, based on modeling standards like TEI (Text Encoding Initiative), OWL (Web Ontology Language), or SKOS (Simple Knowledge Organization System). These standardizations lay the foundation for the semantization and, consequently, the automated processing and analysis of specialist knowledge, incorporating domain-specific ontologies and vocabularies. The separation of content and its presentation as a fundamental feature of XML-based (eXtensible Markup Language) formats implies a high degree of flexibility when dealing with the analysis and transformation of the original (textual) data in different presentation forms, but also calls for standardized workflows in the processing of such data. A variety of solutions and frameworks have been developed for this task. This contribution aims to present and discuss an object-oriented approach to a digital archive, which was implemented in a FEDORA (Flexible Extensible Digital Object Repository Architecture) -based digital repository environment.***

Index Terms: ***Information modeling, Long-term preservation in the humanities, Text Encoding Initiative***

1. Introduction

ISSUES of management and delivery of digital resources are nowadays gaining in importance in museums, archives and libraries, but also in humanities research. In many places, the digitization of collections has begun as an attempt to preserve and considerately make available books, manuscripts and other cultural artifacts. The currently available storage media hardly sets physical limits to mass digitization projects, but there are issues of logistics and sustainability that arise in such projects in many ways.

Thus, as early as 2011, the German Science Council spoke out in favor of the expansion of sustainable, research-suitable digitization and efforts to coordinate the standardization and interconnection of relevant IT-infrastructures.

2. Semantic Enrichment as a Key for a New Methodology

Such research-suitable digitization reaches far beyond a purely pictorial and textual representation of cultural artifacts in the computer. It is not only about digitizing source materials and other scientific resources in the classical sense (colloquially: *scanning them*), i.e. converting them from an analog form of representation into a – in relation to the data format usually proprietary – digital form. Sustainability thereby only arises through the enriched representation of sources, traditions, texts and images aimed at the formal visualization (explication and contextualization) of the semantic structures contained in such data; hence, the enrichment of the content of the digital objects emerging during the process of digitization with (standardized) metadata, referring to different levels of description, such as logical text structure, interpretative or narrative levels, morphology, syntax, etc. Through such enrichment, collections of digitized objects evolve into representations of digital editions: formally enriched cultural artifacts become the empirical data base for research in the humanities by opening up new possibilities for an IT-based representation and analysis of these semantic structures: for example, paleographic properties of a manuscript can be formally differentiated and statistically analyzed, narrative levels or regional references in a literary text corpus can be visualized and opened to interpretation, extensive image collections can be subjected to comparative analyses [9]. The main difference to a paper-based edition is the effort to enrich the edited text with metadata which does not only refer to descriptive (i.e. describing an electronic resource) *data about data*, but also to

Manuscript received April 17, 2014

The Author is with the Center for Information Modeling – Austrian Centre for Digital Humanities, University Graz, Austria. (E-mail: *johannes.stigler@uni-graz.at*)

information which semantically structures and analyses the content of a text document.

Plain-text of a document produced with a word processing application or a transformed PDF file of such a document is from this perspective merely an amorphous mass, whose structuring refers to the lowest unit of characters (spaces, punctuation, etc.). Simple forms of automated enrichment with semantic meaning can be applied at this stage – for example for the recognition of sentence and word boundaries – as well as algorithms which automatically annotate the morphosyntactic properties of words for later use in search processes whose query language allows the search for grammatical structures in a text database.

So-called *repositories* or *digital archives* constitute the necessary IT-infrastructure to keep the data of digital editions sustainable and permanently available. They provide a quotable provision of the available content and thereby organize the persistent survival of information in an ever-changing technological environment. Many national and EU-wide projects and initiatives (CLARIN, Common Language Resources and Technology Infrastructure, http://www.clarin.eu; DARIAH, Digital Research Infrastructure for the Arts and Humanities, http://www.dariah.eu among others) are currently evaluating and developing such systems which are – generally speaking – based on the intention to improve scientific research and communication processes with uniform access to electronic knowledge pools and make them more transparent (to outsiders).

3. Digital Archives as Virtual Research Environments

Digital archives are not merely storage spaces, but also support the processing of digital resources in different scientific scenarios. Flexible authorization models control the web-based access to source materials and research results. Long since, the design of relevant applications is no longer exclusively oriented on the concept of a pure data-store, i.e. an optimized storage and retrieval facility for static content (text, picture, and sound or film documents). In many projects, modeling standards are being defined and workflow models implemented which aim at the digital representation of the entire creation process of scientific research results (e.g. the eSciDoc project at the FIZ-Karlsruhe, https://www.escidoc.org, or Text Grid, http://www.textgrid.de/). Issues of text-, or better, information-immanent annotation and thus the semantization of the content of texts, images, movie clips, etc. come to the fore. Digital data, manually or (semi-)automatically enriched with domain-specific metadata (e.g. text corpora containing lemmata or morphosyntactic information, structurally annotated transcriptions of manuscript, etc.) can not only be researched in an intelligent and ontology-based way but also constitutes important points of reference for empirical analyzes and thus supports the theory construction of the respective scientific domain. This development also renders the classic division of labor between producers and archivists of scientific research results obsolete. Issues of digital archiving or more generally the digital representation and modeling of knowledge are becoming relevant methodological issues in the scientific domain of the respective contents. Hence, in relevant application scenarios, workflows which allow for the collaborative editing and managing of digital resources are to be preferred.

4. The 'Lingua Franca' for Digital Archives

If the *enterprise* digital preservation of scientific and cultural heritage is to prove a sustainable success, it is necessary to establish new avenues for long-term preservation scenarios beyond proprietary data formats: These are too short-lived for sustainable forms of archiving, and some may require error-prone, automated migration processes in the digital archive. Overall, XML-based data formats have prevailed in recent years, not only as a format for descriptive metadata, but also for the overall modeling and annotation of the content of digital objects. As a text format in the humanities, the TEI (Text Encoding Initiative) metadata set offers very flexible and comprehensive options for the human-readable modeling of text documents of almost any origin, based on the premise of separation of content and representation [14]. Overall, it can be observed that the progressive dissemination of XML-based data formats has resulted in the increasing development of XML technologies and tools in applied IT-areas which are also useful for knowledge modeling.

An essential prerequisite for the establishment of XML – far beyond the original application field of (syntactic) information structuring – was the stabilization and consolidation of the standards, especially by (a) the normalization of the structure of well-formed XML documents through the introduction of the XML Information Set standards, (b) the introduction of an extensible type system in XML, which allows the description of arbitrary data structures (XML Schema) and (c) the establishment of a uniform convention for the use of namespaces [4,5]. Thus, the foundation was laid for the development of

complex annotation languages, as represented by the TEI in its present version P5. This version was developed with the intention to construct a universal convention for text annotation and record it in the form of guidelines (applicable to all languages and text types) [3]. These guidelines provide a flexible framework for the definition of (normative) encoding standards whose application areas include sources and documents as they exist or are being produced in a variety of (humanities) disciplines, from historical documents to texts generated in survey situations of quantitative and qualitative empirical social research, to literary and linguistic text corpora [11,12]. In addition to continuous texts, TEI can also be used for the annotation of non-continuous texts such as dictionaries, etc.

Complementary to this primary application of XML-based information modeling, a number of other functional areas in information processing have emerged in which XML plays a central role in digital preservation:

(a) Metadata description and knowledge management, i.e. the secondary use of XML to add additional description elements to content, e.g. by using standards like the Resource Description Framework (RDF, http://www.w3.org/RDF/) [2], Topic Maps (http://www.topicmaps.org/) and Web Ontology Language (OWL, http://www.w3.org/TR/owl-semantics/). These standardizations provide the foundation for the modeling of domain-specific vocabularies as well as making them accessible to automated processing (e.g. in a search process). Extensive examples of such domain-specific ontologies can be found in the field of archival and museum work: the International Committee for Documentation of the International Association of Museums (CIDOC, http://cidoc.ics.forth.gr/) has compiled an extensible ontology for terms and information in the area of cultural heritage in the CIDOC Conceptual Reference Model. The SKOS specification (Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/) presents a first-time effort by the W3C to standardize a formal language for encoding documentation-languages such as thesauri, vocabularies and other controlled vocabularies based on the Resource Description Framework.

(b) Transformation of information, i.e. the use of XML standards to map information structures to each other, e.g. to derive a representation format from an XML structure. Here, Extensible Stylesheet Language (XSL, http://www.w3.org/TR/xslt20/) affords special consideration as the umbrella term for a complex control system, which consists of three specifications: (1) XSLT (XSL Transformation), a transformation language for the structural editing of XML documents which describe a rule-based transformation process from an input file into one or more output files of any target format using XML syntax, (2) XPATH (XML Path Language), which allows the selection of (virtual) sub-trees of an XML tree structure, and (3) XSL:FO (XSL Formatting Objects), a standard for printed page description.

(c) Exchange of information, which is the use of XML as a universal data exchange format between applications, also at the level of protocols over the internet.

(d) Application modeling language, that is the use of XML in the design, programming and deployment of applications, e.g. in UML-based (Unified Modeling Language) development environments, but also as a control-relevant modeling language for process sequences in web-based application frameworks.

5. An XML-based Publication Framework

For the processing of XML-based data, a number of specific publication frameworks exist: They support, among other things, the representation of different aspects of a multi-layered, modeled text structure in different formats (or views) for scientific end-users. In general, they are characterized by the following features: (a) realization as a client-server application based on W3C standards; (b) use of a Three-Tier- structure, where the client is usually a web browser and a corresponding server-sided application logic is implemented; (c) modular integration of XML-processing components (XSLT processor) and (d) separation of the (XML) content from its presentation aspects to achieve flexibility with respect to the output generation and processing of the data.

As an open source project that implements this requirement profile, the COCOON framework (http://cocoon.apache.org) developed by the Apache Software Foundation has gained wide acceptance. COCOON is a Java-based, dynamic component framework that can be integrated into a JSP (Java Server Pages) container environment and implements certain standard components alongside its own modules for controlling XML-processing. The open, component-based architecture allows for easy integration of database- or authorization-modules. Like almost any web application, COCOON is embedded in a fixed request-response-cycle. A request in COCOON is implemented as a pipeline of successive steps. Within COCOON, the URI of a request is

analyzed and passed on for processing to one of the pipelines defined in the central control file through a so-called Matcher. Each step of the procedure can write data depending on the current pipeline state and read from various data sources (file, [XML-]database, web service, etc.). With this simple approach, complex tasks can be disassembled into several subtasks. Communication within a pipeline occurs via so-called SAX (Simple API of XML) streams, in which responses of a pipeline component are passed on to the next component in the processing chain in XML format. In the last step of the pipeline, the output which is transmitted back to the client is generated (where again, XML components can take over the serialization of the XML stream). Standard components of the framework are thus well suited for the realization of (multilingual) web applications with underlying XML-based data and text bases.

6. Asset Management Systems

So-called asset management systems offer a framework for digital archives going far beyond the functionality of COCOON [1]. The term *asset* in this context describes the smallest unit which is structuring, described, and managed by the system, comparable to a catalog entry. Such an asset is made up of a primary data stream (e.g. text document, spreadsheet, presentation file, audio or video file etc.) and at least one set of descriptive metadata (Dublin Core). Such so-called *simple model assets* stand opposed to *compound model assets* which can consist of a wide range of primary data streams and associated functions: e.g. an asset for digital books, consisting of all image files of photographs of a manuscript, the edited text (in any text format) and a set of methods which allow scrolling through the pages of this (virtual) book or zooming details of the individual pages. They are used quite generally for the storage and management of digital resources, as they occur in scientific contexts in great variety. In contrast to content management, asset management puts special emphasis on aspects of sustainable, metadata-based and quotable archiving of and access to digital resources, controlled through flexible permission models. Sustainability in this context refers to the long-term availability of resources, but also aims at fundamental considerations in connection with archiving projects, such as recommended (and actual) data formats. In the context of scientific data management, it is paramount to ensure that – regardless of changing software environments – texts, pictures, movies, statistical data bases and other materials archived for scientific purposes remain quotable and safely accessible over longer periods of time (comparable to print publications) [15].

As basic functions to accomplish its tasks, an asset management system provides import and export operations for data sources (possibly connected with format conversions, such as from MS Office to XML), the option of enrichment with different descriptive but also administrative metadata, the pooling of resources in containers and the versioning of data sources, as well as strategies for the URL-based addressing of individual sub-entities of an asset.

7. Flexible Extensible Digital Object Depository Architecture

Furthermore, an asset management system as described here is geared towards realizing the paradigm of Single Source Publishing: content can occur in a variety of representation formats which are dynamically generated (when referencing the content through a web-browser) from the content stored in an asset using so-called style sheets. When searching for a suitable platform for the realization of a suitable IT-infrastructure for the long-term preservation and archiving of digital (research) data at the Center for Information Modeling – Austrian Centre for Digital Humanities, extensive research ultimately led us to the open source project Flexible Extensible Digital Object Repository Architecture at Cornell University (FEDORA, http://www.fedora-commons.org).

Our own project with the acronym GAMS (Geisteswissenschaftliches Asset Management System, http://gams.uni-graz.at) is a digital archive for the metadata-based management and sustainable provision of digital resources. From manuscript to video, from text edition to image archive, it offers to faculty and students the means to archive and publish such resources in a standardized, quotable and web-based way. Heterogeneous requirement profiles are indicative of application scenarios of digital archives in scientific contexts: the specific areas of operation of the system range from learning object collections to digital editions, from video and film archives to (morphosyntactically) annotated and multimodal corpora. GAMS is also integrated in several European research infrastructure projects and is therefore not a mere island solution. As an example, the content managed through GAMS is integrated in Europeana (http://europeana.eu), a long-term EU project with the aim of initiating and implementing a common European search portal for scientific content.

On a structural level, the process of digitization also requires further thinking on sustainability. The European research initiative DARIAH aims

at the establishment of a sustainable digital research infrastructure in the university environment throughout Europe. Based on such research infrastructures, the sharing of resources, methods, data and experiences will be encouraged and scientists will be supported in establishing collaborative and digital research cultures to respond to their genuine research questions in new ways and even develop new research questions. Strategically, these projects aim at joint software development as well as the construction of competence centers to create the necessary institutional prerequisite for the further digitization of humanities.

At its core, FEDORA provides a database-driven, modular expandable storage and management structure (repository) for arbitrary (distributed) digital resources, with web-based access, guided by the principles of a Service Oriented Architecture (SOA) with the following properties [10]:

(a) Web-based (SOAP, Simple Object Access Protocol, http://www.w3.org/TR/soap/), platform-independent, distributed system architecture,

(b) Apache Lucene-based full-text index and versioning management of the asset contents, (http://lucene.apache.org/)

(c) RDF-based triplestore with the SQL-like query languages ITQL (Mulgara Project, http://docs.mulgara.org/tutorial/itql.html) and SPARQL (SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/).

(d) Definition of intricately controllable access rights to assets and their sub-entities based on XACML (Extensible Access Control Markup Language, http://www.oasis-open.org)

(e) Standards-based import and export formats: METS (Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/) and others,

(f) Unique URL-based addressing of digital resources,

(g) Support of standardized protocols for metadata exchange, like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and others (http://www.openarchives.org/),

(h) The option to realize system environments with corresponding numbers of concurrent users by repository clustering and load balancing thanks to the carrier technology (Apache Tomcat, http://tomcat.apache.org/) [9].

As one of the most important paradigms in the field of software design, Wolf – referring to Gamma and Pree [16, 13] – identifies the formation and implementation of the concept of object orientation, which, in concert with the introduction of modular software concepts with high granularity and the systematization and standardization of object- and component-oriented software development through design patterns, has led to a high degree of reusability of software technologies.

Object orientation is characterized, among others, by the following properties:

(a) The definition of classes with associated properties (attributes) and methods as essential structural elements and

(b) The formation of class hierarchies through the principle of inheritance, taking advantage of polymorphism.

These principles are implemented in FEDORA not only at the level of system development, they also structure the application logic at the user level: Through the design of content models (object classes), complex object class hierarchies can be constructed in a FEDORA-based data repository [7,8]. Content models describe not only the content structure for an asset class (data streams) and the potential relations to other objects (container assets), but can tie so-called disseminators (methods) to the data of an asset through the use of Web Service Description Language (WSDL, http://www.w3.org/TR/wsdl.html): e.g. XSLT transformations which convert the XML data streams of an asset to any desired target formats (HTML, PDF, etc.); methods that convert a color image stored in an asset, into a black-and-white variant for the use of the image in offset printing; functionalities allowing the navigation in an indexed video file, and other features. With respect to an object model specifically geared to modeled text corpora, which is initialized at its instantiation with an XML file encoded in TEI format, these could be dissemination methods which in one instance represent an XML data stream as a navigable HTML document (e.g. with specific interactive analysis options, like the configurable highlighting of certain text structure levels), and in another instance as a PDF or LaTeX file.

FEDORA also supports the assignment of intricately configurable access rights to assets and their dissemination methods. Basically, all entities which are part of an asset can be individually addressed and referenced (web-based). Through XACML, individual access channels can be linked to distinct authentication and authorization rules. For example, all access paths to a text object except that which produces the representation in HTML format can be restricted to authorized project staff. FEDORA also supports standards such as LDAP (Lightweight Directory Access Protocol, http://www.openldap.org), Shibboleth (http://shibboleth.internet2.edu/).

Through this feature, related object data can

be stored in a common administrative and storage context (asset) while at the same time implementing query and editing scenarios controlled through differentiated access models.

8. CONCLUSION

Standardized annotation languages and technologies for the processing of XML-based data structures based on relevant reference models nowadays form the basis for the realization of sustainable long-term preservation and archiving scenarios for research data in the humanities. In such contexts it is a requirement to provide distributed digital resources through centralized storage, management and retrieval structures and thus ensure the quotable archiving of digital knowledge bases on the premise of recyclability. Although much has already been accomplished to that end, most of the existing technical solutions are essentially prototypes with a low degree of standardization. Many technical issues have been resolved, yet several desiderata remain:

(a) The development of standardized modeling languages for the description of processing workflows and forms of representation,

(b) An open methodological discourse on issues of the digital transmedialization of research data in the humanities,

(c) Adequate and standardized tools and corresponding technical infrastructures

(d) *Trusted* digital archives and sustainable institutional infrastructures.

REFERENCES

[1] Austerberry, D., “Digital Asset Management. How to realize the value of video and image libraries,” *Focal Press*, 2004;

[2] Brickley, D., Guha, R. V. (eds), “RDF Schema 1.1. W3C Recommendation, 25 February 2014,” http://www.w3.org/TR/rdf-schema, 12.05.2014;

[3] Burnard, L., “What is the Text Encoding Initiative?,” *OpenEdition Press*, 2014.

[4] Cagle, K., Duckett, J., Griffin, O. et al., “Professional XML Schemas,” *Wrox Press*, 2001;

[5] Eckstein, R., Eckstein, S., “XML und Datenmodellierung : XML-Schema und RDF zur Modellierung von Daten und Metadaten einsetzen,” *Dpunkt-Verlag*, 2004;

[6] Gamma, E., Helm, R., Johnson, R. et al., “Entwurfsmuster : Elemente wiederverwendbarer objektorientierter Software," *Addison-Wesley*, 2007;

[7] Green, R., “University of Hull digital colour image object specification,” http://www.hull.ac.uk/esig/repomman/downloads/INT-D3-1-imageObject-v03.pdf, 12.05.2014;

[8] Green, R., “University of Hull digital public document object specification,” http://www.hull.ac.uk/esig/repomman/downloads/INT-D3-3-documentObject-v01.pdf, 12.05.2014;

[9] Hofmeister, W., Stigler, H. J., “Die Edition als Interface. Möglichkeiten der Semantisierung und Kontextualisierung von domänenspezifischem Fachwissen in einem Digitalen Archiv am Beispiel der XML-basierten ‚Augenfassung‘ zur Hugo von Montfort-Edition,” *De Gruyter, Internationales Jahrbuch für Editionswissenschaft* 24/1, 2011, pp.79-95.;

[10] Lagoze, D., Payette, S., Shin, E. et al., “Fedora. An Architecture for Complex Objects and their Relationship,” http://www.arxiv.org/ftp/cs/papers/0501/0501012.pdf, 12.05.2014;

[11] Lobin, H., “Erweiterte Dokumentgrammatiken als Grundlage innovativer XML-Tools,” *De Gruyter, Information Technology* 45/3, 2003, pp. 143-150.;

[12] Lobin, H., “Textauszeichnung und Dokumentengrammatiken,” *Texttechnologie. Perspektiven und Anwendungen,* Stauffenburg, 2004, pp. 51-82.;

[13] Pree, W., “Komponenten-basierte Softwareentwicklung mit Frameworks,” *dpunkt-Verlag*, 1997;

[14] TEI Consortium, “TEI P5: Guidelines for Electronic Text Encoding and Interchange, 20th January 2014,” http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf, 12.05.2014;

[15] Witten, I. H., Bainbridge, D., “How to build a Digital Library,” *Morgan Kaufmann*, 2003;

[16] Wolf, C., “Systemarchitekturen. Aufbau texttechnologischer Anwendungen,“ *Texttechnologie. Perspektiven und Anwendungen,* Stauffenburg, 2004, pp. 166-192.