MASARYKOVA UNIVERZITA

FILOZOFICKÁ FAKULTA

KATEDRA INFORMAČNÍCH STUDIÍ A KNIHOVNICTVÍ

Thesis Project Proposal

User-Centered Conceptual Models for Understanding Web Archive Collections

Master's thesis

Bc. Illyria Brejchová
525117

Supervisor:
PhDr. Michal Lorenz, Ph.D.

Czech title: Na uživatele orientované konceptuální modely pro pochopení dat z tematických sbírek webových archivů

1. Statement of the Research Problem

Since the rise of the World Wide Web in the 90s, web-based works have become increasingly prevalent, often replacing the role of more traditional print resources. The informational, historic and cultural value of digital born documents on the web is undeniable, yet these resources are highly ephemeral. Over the past 25 years, countless web archive initiatives have been undertaken around the globe by various types of institutions and in varying scope. Extensive collections now exist, yet they are not without challenges regarding the long-term preservation of the digital objects, their metadata description, and usability.

Research has shown that current metadata practices are not sufficient relative to the needs of the designated community (Costea, 2018; Truman, 2016; Venlet et al., 2018). While institutional metadata handbooks exist, they vary significantly. A guideline for web archive metadata description published in 2018 seeks to bridge these institutional practices and takes a user-centered approach (Dooley & Bowers, 2018). However, they do not make their recommendations based on any domain specific conceptual model nor was conceptual modeling in their scope. And while existing bibliographic and archival data models are designed to be extensible to any kind of resource, archived websites have many specific characteristics that should be taken into consideration, such as relationships between the live and archived website, between different versions of the same website, between different websites sharing the same URL in time or between contributors, domain owners and copyright holders.

While it is not feasible to describe archived web resources with such extreme granularity, it is important to understand these relationships at least in theory. With solid conceptual modeling, it will be clearer what is being described. Web archive metadata specialists can then make better decisions about how to represent significant properties of the archived resources in metadata formats and information systems. Such a conceptual model would also benefit researchers seeking a deep understanding of the archived web resources they are studying.

2. Overview of the Literature, Research Context, and Relevant Projects

2.1 User Needs

Truman (2016) studied the situation in 23 web archives looking at their characteristics and which barriers researchers encounter when using them. They identified 22 opportunities for researchers in four broad categories: 1) The largest category is communication and collaboration, here they emphasize the need for closer cooperation between researchers and curators. Curators must better document and share their archival policy and how it and other technical parameters affect the archived collection. This documentation must be accurate, which was not always the case, as the study found. Researchers also require assistance from web archivists with data extraction, and curators need input from researchers regarding their methodological requirements for the data and metadata for the archived content to be analyzed in an academically rigorous way. 2) The study stresses the need for developing standardized and interoperable tools for processing big data from web archives and iteratively expanding them to meet researchers' needs. A need for better discoverability of web archived content both within each institution and across other web archives was also identified. 3) Education and skill development are important for employees of web archives as well as scholars, who seek to study data in a new and unfamiliar format. There is also the potential for educating web developers and working toward a web that is technically archivable. 4) There is a need for expanding the capacity of web archives in respect to both staff and storage space.

Maria-Dorina Costea (2018) published a report of the usability of web archive content for research based on a survey and interviews with Danish Humanities researchers. Attention was paid to both the Internet Archive and the Danish web archive Netarkivet. Both users and potential users of web archived data were included, with the majority of respondents having no prior experience with web archiving. Of the researchers that had experience with using archived web content in their work, most did so for qualitative research in history or analysis of political discourse online. Only one respondent worked on a quantitative study, however, they were unable to complete it due to technical problems. The researchers expressed a need for better search tools, such as full text, image or audio recording search capabilities, the option to export data into a standard format and an integrated workspace for academic work. They emphasized the need for more detailed documentation that would make the collection practices and tools more transparent. They would appreciate more metadata and for the existing metadata to be more accurate. If part of the archived website was unable to be archived or is displayed differently than it would have been on the live web, they need to know about it. They would also like the functionality of being redirected to a better captured instance of the archived web in such situations. The researchers lacked the functionality to filter search results and would appreciate a more complete and easier to understand tutorial as well as explanations to domain specific vocabulary.

OCLC published a literature review of user needs conducted as part of their efforts to formulate recommendations of descriptive metadata for web archives. The research focuses largely on the needs of academic researchers. It was found that potential users still lack an understanding of what web archives are and how to use them. Those that do use them express a strong need for provenance metadata and transparency regarding the acquisition of the archived websites. Users generally lack the technical knowledge necessary to interpret and use web archive data, resulting in the need for user friendly tools and interfaces, better user support, and outreach. As for web archive practitioners, they need scalable descriptive practices, hybrid bibliographic and archival approaches to cataloguing, and interoperable metadata across systems (Venlet et al., 2018).

2.2 Conceptual Data Models

LRM (Library Reference Model) is the most recent high-level conceptual reference model of the bibliographic universe published by IFLA (The International Federation of Library Associations). It consolidates FRBR (Functional Requirements for Bibliographic Records), FRAD (Functional requirements for Authority Data) and FRSAD (Functional Requirements for Subject Authority Data). LRM is an entity relation model and defines 11 entities at three hierarchical levels: Res, Work, Expression, Manifestation, Item, Agent, Person, Collective Agent, Nomen, Place and Time-span. (Riva et al., 2017) While this model addresses many of the limitations of FRBR and should in theory be applicable for modeling the entirety of the bibliographic universe, which as of now includes web archived resources, it remains highly centered on traditionally published recourses. The concept of publication in the context of online content is problematic, for this reason conceptual models from the museum and archiving tradition could be better applicable. The most significant of these is the extensible object-oriented Conceptual Reference Model (CRM) of CIDOC (International Council of Museums). CRM is meant to function as sematic “glue” for disparate cultural heritage datasets by providing a formal structure for describing concepts and relationships. Attempts at bridging the bibliographic and archival traditions have already been made, starting with object oriented FRBR (FRBRoo), the first version of which was published in 2006. Most recently work on LRMoo is being conducted (*LRMoo*, 2021). The aim of these conceptual models is to enable semantic interoperability. A conceptual model of the web archiving universe of discourse should therefore be compatible with these models already in place and so I will be drawing upon them significantly.

In 2018 the OCLC Research Library Partnership Web Archiving Metadata Working Group published detailed user-centered recommendations for descriptive metadata for web archives. It was found that the application of standards for descriptive practices are highly inconsistent, a combination of bibliographic, archival and hybrid approaches are in use. Moreover, the descriptive guidelines that are in use do not usually take into consideration unique characteristics of either live or archived web content. The most used standards are Resource Description and Access (RDA), Describing Archives: A Content Standard (DACS) and Dublin Core. The working group proposes the use of 14 data elements (Collector, Contributor, Creator, Date, Description, Extent, Genre/Form, Language, Relation, Rights, Source of Description, Subject, Title, URL) and give descriptions and examples for their use (Dooley & Bowers, 2018).

Some of the most recent direct research into implementing conceptual models for web archives derives from the internet art preservation community. Rhizome, a New York based non-profit dedicated to digital preservation, exhibition and scholarship of born-digital art, have pilot tested the implementation of PROV-DM for describing provenance metadata of internet art works in ArtBase, their digital art archive. They realized that the metadata available to their users, especially that relating to provenance, was insufficient for contemporary art historians. They decided to implement PROV-DM because it best enables the representation of the lifecycle internet art undergoes, the various actors who contribute to its creation, development over time and preservation, as well as the relationships between variations of the artwork both on the live web and in the collections of cultural institutions. Just as importantly, it allows the representation of unknown data and the description of objects at varying levels of detail. In this case study, they also used PROV-O to map PROV-DM to RDF in a linked data knowledge management system. The metadata model tested is also compatible with other bibliographic semantic models such as CIDOC-CRM and FRBRoo used in other cultural heritage systems (Rossenova et al., 2019). Those had, however, been overwhelmingly developed for the description of primarily physical documents and are more narrowly conceived. On the other hand, metadata schemas developed for digital documents such as PREMIS do concentrate on describing processes, but are too abstracted to be useful in the case of internet art (Rossenova, 2021).

Bahry et al. (2022) developed a conceptual data model for the Malaysia public figures web archive as part of the database design process for a new digital repository. The model was visualized in the form of an ER Diagram, therefore concentrating on identifying and describing entities, attributes and relationships. Existing web archiving metadata standards were benchmarked and the most suitable were mapped to the conceptual data model and included as data elements. The resulting data model is designed to be easily understandable to novice web archivists, researchers and student users alike and can be adapted into real metadata modelling for the web archive repository. This conceptual data model was created for a more specialized web archive than I have in mind for my thesis, even so, it can serve as a good foundation to be expanded upon for the purposes of my own research on this topic.

Lieber et al. (2021) from the Royal Library of Belgium describe how the Europeana Data Model (EDM) in combination with PREMIS and PROV can be used to represent social media collections in an interoperable way. They take a Competency Question based approach, to identify and express requirements for the subsequent ontology engineering. This involved gathering and identifying user stories to decide which metadata fields are most significant. The solution they came to involved creating a semantically described RDF Knowledge Graph. This allows for automated metadata description at the level of individual posts while also retaining context information from the harvest as a whole. There is also the possibility of adding or editing metadata manually. Specific standardized metadata records in MARC or EDM can then be generated from the Knowledge Graph implemented as a relational database with no added manual labor. Aggregated data from archived posts can also be generated at the collection level. A big benefit of this approach is that the metadata can be made publicly available, while the archived items are not, due to copyright. While this paper emphasizes the application side of the developed model rather than describing the developed ontology in detail, this work is still highly relevant to my own conceptual modeling ambitions.

2.3 Theses

As for recent Czech theses dealing with web archiving, I could mention the bachelor’s thesis of Ondřej Kadlec (2020), where he maps the technical workflow of the Czech web archive (Webarchiv) of the Czech National Library and the tools used. Of interest is his discussion of the mapping of current processes of Webarchiv to the reference model OAIS, an ISO standard stating the requirements an open archival system should have for securing the long-term preservation of digital data. It was found that not all requirements of the OAIS model are fully fulfilled. Namely, the archival information packets lack semantic and interpretation metadata and Webarchiv lacks an entity responsible for the long-term preservation and understandability of the archived data by the designated community. Jaroslav Kvasnica (2016) also researched the long-term digital preservation resources in Webarchiv in depth in his masters’ thesis with an emphasis on OAIS and metadata. This thesis is relevant for its detailed description of how metadata is created and stored in Webarchiv. Lastly, I consider the master's thesis of Jan Vokřál (2021) to be relevant for his use of conceptual modeling, though not of the web archiving universe of discourse, but rather in that of fiction.

3. Motivation for and Impact of the Research

I worked part-time for a year as a curator at Webarchiv of the National Library of the Czech Republic and found the work to be very meaningful. Preserving Czech digital national heritage is important and a great responsibility, yet the character of the resources, their technical properties, and the needs of the users change rapidly. Therefore, the research and development into web archiving always lags behind that of the live web. As pessimistic as that sounds, I choose to view it as a challenge and opportunity. Over the past two years I have found the web archiving community to be dedicated, adaptable and inspiring and I hope to contribute to the efforts of this community of practice and research with my own findings.

During my work for Webarchiv, which also involved creating descriptive metadata for the archived websites, I felt a need for a better understanding of the designated community and user needs, as well as of the significant properties of the resources and the relationships between them. This is a need felt by the wider web archiving community as well, as is clear from the literature review, as well as from talking to employees of the web archive of the Dutch National library, where I am applying for an internship.

A good conceptual model will help web archivists and researchers alike better understand the contents of web archives. Well-defined entities will make it clearer what is being described by the available metadata and enlighten relationships between resources. It will also assist web archive metadata specialists in making better decisions regarding which descriptors are significant relative to their designated community. And last but not least, it may serve as a solid foundation for designing data structures for specialized information systems dealing with web archived content.

4. Research Questions, Aims and Goals

The aim of my research is to create a high-level conceptual model describing the web archiving universe. The modal shall be grounded in a deep understanding of user requirements, representative of current domain knowledge, and compatible with CRM and FRBRoo/LRMoo but implementation agnostic. The goal of the research is to provide a shared framework for understanding and describing web archived resources. This should benefit web archivists when managing and creating metadata, developers when designing and implementing web archive information systems, and academic users seeking a theoretical model for understanding archived web resources. My research question is therefore: To what extent can the web archiving universe of discourse be represented in existing conceptual models from related fields?

5. Research Design

Drobíková et al. (2018) describe the use of conceptual modeling within LIS. A conceptual model is an explicit representation of a domain conceptualization using a modeling language. For the purpose of this work, I will use the Unified Modeling Language (UML). UML is a widely used standardized general-purpose object-oriented modeling language (Object Management Group, 2017). It is suitable for this work since the conceptual model of the web archiving universe will be developed from existing object-oriented conceptual models in the realms of cultural heritage (CIDOC CRM), and librarianship, (FRBRoo and LRMoo). The modeling language is used to visualize semantic relations between objects and entities.

The model being created is to be a high level descriptive conceptual reference model of the web archiving universe of discourse. Kučerová (2018) describes the process of creating a conceptual model and grounds it in the three world and objective and subjective knowledge philosophy of Karl Popper. The first phase of creating a conceptual model involves creating a subjective mental model of the system. This is something we do implicitly as one makes sense of the world, though it can also be done purposefully with the intent of communicating the mental model. The second phase involves creating an objective model. First the subjective model is expressed using a language and then it is manifested on a medium.

A systems analysis approach will be taken to the design of the conceptual model. The web archiving universe of discourse will therefore be conceived of as a complex open system. Given web archiving systems are already well established, a bottom-up approach will be taken, and the abstract conceptual model will be in part reverse engineered from an existing understanding of the system. First the system requirements will be defined based on user practices and then the model will be created through an iterative process of decomposing the system (analysis) and abstracting it (synthesis). Through analysis of the system relevant objects, classes and entities are identified and through synthesis hierarchical relations between the objects are identified (Kučerová, 2017).

To aid in the system analysis, selected stakeholders with a deep understanding of web archiving will be interviewed. During the interview, the method of card sorting will be employed to gain insight into their own implicit mental models of the web archiving system. I estimate needing to talk to at least six stakeholders, ideally three web archiving professionals and three users. The results will be analyzed, and insights will be integrated into the final model.

The comparative method will be used to compare the conceptual model of the web archiving universe to conceptual models in related fields and to identify whether there are concepts within web archiving that do not map to them.

6. Schedule

SUMMER 2022: I will be attending an ERASMUS+ internship at the National Library of the Netherlands, where I shall be creating a report with recommendations for the library concerning the management of metadata of web archived resources. While this is not part of my thesis research, it is thematically closely related. I hope to gain insight into current web archive metadata practices and concerns outside the Czech Republic. I will also have a chance to expand and apply my current domain knowledge with expert guidance.

FALL 2022: Write the theoretical background of the thesis – a brief description of the historic evolution of web archiving and selected web archives, an overview of relevant existing data models, current metadata formats and practices, and user needs. Identify key stakeholders and interview them to gather knowledge relating to the system potentially missing from the literature review.

WINTER 2023: Prepare questions and concepts for the card sorting based on my current understanding of the system. Identify key stakeholders and interview them using the method of card sorting.

SPRING 2023: Conceptual modeling of the web archiving universe of discourse grounded in a system analysis of the web archiving domain and user needs. Documentation explaining the identified objects and relationships between them. Comparison of the web archiving conceptual model to those in related fields.

SUMMER 2023: Mapping of the conceptual model to a selected thematic collection of web archived resources to demonstrate and discuss its expressive power and weaknesses.

FALL 2023: Final revisions.

7. Required Resources

My research will require access to the university's academic electronic resources and to publicly accessible web archive catalogues. I will also need access to a suitable modeling software. Thankfully, there are many freely available open-source options, such as Diagrams.net, as well as proprietary online tools with student licenses, like Vertabelo. I am also able to consult my work as needed with web archive specialists at the National Library of the Czech Republic and at the National Library of the Netherlands thanks to pre-existing professional connections.

8. References

Bekiar, C., Doerr, M., Le Boeuf, P., & Riva, P. (Eds.). (2021). *LRMoo (formerly FRBRoo) object-oriented definition and mapping from IFLA LRM (version 0.7)* (p. 62) [Draft]. https://www.cidoc-crm.org/frbroo/sites/default/files/LRMoo_V0.7%28draft%202021-06-29%29.pdf

Costea, M. (2018). *Report on the scholarly use of web archives*. NetLab.

Dooley, J., & Bowers, K. (2018). *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. 58. https://doi.org/10.25333/C3005C

Drobíková, B., Římanová, R., Souček, J., & Souček, M. (2018). *Teoretická východiska informační vědy: Využití konceptuálního modelování v informační vědě* (Vydání první). Univerzita Karlova, nakladatelství Karolinum.

Kadlec, O. (2020). *Popis technického workflow Webarchivu NK ČR a užívaných nástrojů* [Bakalářská práce, Masarykova univerzita, Filozofická fakulta]. https://is.muni.cz/th/n1dx9/BP.pdf?info=1

Kučerová, H. (2017). *Organizace Znalostí*. Karolinum Press. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5720124

Kučerová, H. (2018). *Pojem modelu a pojmový model v informační vědě*. *29*(2), 5–32. https://knihovnarevue.nkp.cz/archiv/2018-2/recenzovane-prispevky/pojem-modelu-a-pojmovy-model-v-informacni-vede#:~:text=Pojmov%C3%A9%20modely%20v%20informa%C4%8Dn%C3%AD%20v%C4%9Bd%C4%9B,sch%C3%A9mata%20a%20slovn%C3%ADky%20hodnot%20metadat.

Kvasnica, J. (2016). *Dlouhodobé uchování webového obsahu* [Magisterská diplomová práce, Karlova Univerzita, Filozofická fakulta].

https://dspace.cuni.cz/bitstream/handle/20.500.11956/82967/DPTX_2012_1_11210_0_345026_0_129352.pdf?sequence=1&isAllowed=y

Lieber, S., Van Assche, D., Chambers, S., Messens, F., Geeraert, F., Birkholz, J. M., & Dimou, A. (2021). BESOCIAL: A sustainable knowledge graph-based workflow for social media archiving. *Further with Knowledge Graphs : Proceedings of the 17th International Conference on Semantic Systems*, *53*, 198–212. https://doi.org/10.3233/ssw210045

Object Management Group. (2017). *OMG Unified Modeling Language (OMG UML) version 2.5.1* (p. 796). Object Management Group. https/www.omg.org/spec/UML/

Riva, P., Le Boeuf, P., & Žumer, M. (2017). *IFLA Library Reference Model A Conceptual Model for Bibliographic Information* (p. 101). IFLA. https://repository.ifla.org/bitstream/123456789/40/1/ifla-lrm-august-2017_rev201712.pdf

Rossenova, L. (2021, June 3). *Modeling Net Art Provenance: A New Approach To The Interface of Rhizome's ArtBase Archive*. YouTube. https://www.youtube.com/watch?v=6QUUM2tQxmQ&ab_channel=SAKIPSABANCIMÜZESİ

Rossenova, L., Espenschied, D., & de Wild, K. (2019). *Provenance for Internet art: Using the W3C PROV data model*. 16th International Conference on Digital Preservation, Amsterdam. https://ipres2019.org/static/proceedings/iPRES2019.pdf

Saiful Bahry, F. D., Amran, N., Putri, T. E., & Ramli, M. I. (2022). Database design of the Malaysia public figures web archive repository: A social and cultural heritage web collections. *Collection and Curation*. https://doi.org/10.1108/CC-09-2021-0025

Truman, G. (2016). *Web Archiving Environmental Scan* (Harvard Library Report.). https://dash.harvard.edu/handle/1/25658314

Venlet, J., Farrell, K. S., Kim, T., O'Dell, A. J., & Dooley, J. (2018). *Descriptive Metadata for Web Archiving: Literature Review of User Needs*. https://doi.org/10.25333/C33P7Z

Vokřál, J. (2021). *Informační hodnota beletrie: Konceptuální model pro popis beletrie v informačních systémech* [Masarykova univerzita, Filozofická fakulta]. https://is.muni.cz/auth/th/nspln/Informacni_hodnota_beletrie.pdf

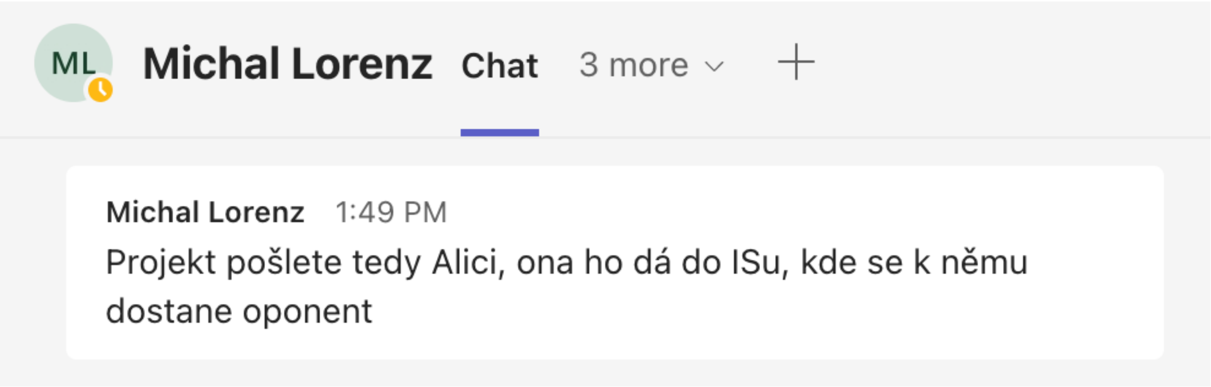
9. Profiling

I aim to profile my studies in Information and Data Management, which is also reflected in my thesis. Conceptual data models are key for designing, implementing and managing metadata in information systems in a way which is functional for the designated community of an information system. This is especially true for web archives which preserve especially complex digital objects.

10. List of Acronyms

CIDOC-CRM	Conceptual Reference Model of the International Council of Museums
DACS	Describing Archives: A Content Standard
EDM	Europeana Data Model
ER Diagram	Entity Relationship Diagram
FRAD	Functional Requirements for Authority Data
FRBR	Functional Requirements for Bibliographic Records
FRBRoo	Object Oriented Functional Requirements for Bibliographic Records
FRSAD	Functional Requirements for Subject Authority Data
IFLA	The International Federation of Library Associations
ISO	International Organization for Standardization
LIS	Library and Information science
LRM	Library reference model
LRMoo	Object Oriented Library Reference Model
MARC	Machine-readable Cataloging
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
PREMIS	Data Dictionary for Preservation Metadata
PROV-DM	Provenance Data Model
PROV-O	Provenance Ontology
RDA	Resource Description and Access
RDF	Resource Description Framework
UML	Uniform Modeling Language

This thesis project has been approved by my thesis supervisor.



Michal Lorenz 1:49 PM

Projekt pošlete tedy Alici, ona ho dá do ISu, kde se k němu dostane oponent