

Towards Supporting Context-oriented Information Retrieval in a Scientific-Archive based Information Lifecycle

Felix Engel¹, Claus-Peter Klas¹, Holger Brocks¹, Alfred Kranstedt², Gerald Jäschke³,
Matthias Hemmje¹

Address¹: FernUniversität in Hagen, Universitätsstr. 1, 58097 Hagen, Germany

Email¹: Holger.Brocks, Felix.Engel, Claus-Peter.Klas, Matthias.Hemmje@FernUni-Hagen.de

Address²: Deutsche Nationalbibliothek, Adickesallee 1, 60322 Frankfurt am Main, Germany

Email²: A.Kranstedt@d-nb.de

Address³: GLOBIT GmbH, Julius-Reiber-Str. 15a, 64293 Darmstadt, Germany

Email³: gerald.jaeschke@globit.com

Abstract

Supporting access to archived scientific publications, supplementary data, and multimedia objects as a basis for various types of reuse in scientific work processes and in publication processes is still an open issue in many ways. Reuse comprises, for instance, the subsequent verification of the content or its exploitation with a novel purpose. Retrieval approaches that factor in the versatile context of the archived data and documents can contribute to supporting reuse beyond traditional indexed based retrieval. The capturing of additional metadata during all life phases of digital objects before, during and after archival is a prerequisite to this approach. This paper motivates the usage of captured context data of digital objects for the purpose of enabling efficient reuse of preserved digital objects.

Keywords: OAIS, IR, context, scientific publishing

Introduction

An important goal of Digital Preservation (DP) is to enable the reuse of digital content. Reuse of digital content covers its subsequent verification and its exploitation with a novel purpose. Understanding the nature of the digital content and its origin supports information seekers in identifying relevant elements in archive collections and in interpreting them correctly. But the preservation of digital content, especially in the long term, covers periods of time, during which the nature of digital resources as well as their usage settings change [10]. As consumers cannot refer back to the creators, reuse of preserved digital objects depends on proper descriptions provided through the archive.

The SHAMAN (*Sustaining Heritage Access through Multivalent ArchiviNg*) project, co-funded by the European Commission under the seventh RTD Framework Programme, aims to develop a next generation digital preservation framework. The context model developed within the project provides an infrastructure-independent representation of the attributes associated with and (implied) relations between digital objects. Provided that the archive manages, preserves and makes available context data about digital objects, the SHAMAN context model is a potentially invaluable source for context-oriented retrieval on the archive holdings.

For SHAMAN context is not only defined by the discrete digital objects themselves, but also by the processes, in which they were *created*, *ingested*, *accessed* and *reused*. Processes are

organized along phases within an *Information Life Cycle Model*. From an archive-centric perspective, each phase identifies one distinct stage in the life cycle of digital objects.

Context comprises information about the preserved object itself, but also the relations between objects. Hence, capturing of contextual data is of great interest for enabling advanced retrieval in archival access, in addition to supporting preservation actions. Through the preparation of context data the retrieval is not restricted on full text index, but could be opened to retrieval approaches on relations between objects. The retrieval results in this case are not necessarily preserved objects but could also be sets of contextual data.

This paper motivates the approach of context oriented Information Retrieval (IR), in an archive based life cycle with a focus on scientific publishing. This comprises current context oriented approaches in IR as well as the current approach towards a Context Model and an Information Life Cycle Phases Model in the *SHAMAN* project.

1. Contextual Data in the Domain of Scientific Publishing

Today, scientific publications are expected to be by origin born digital. They are presented and discussed at conferences and preserved over time in archives. Conferences take a prominent part in scientific research, because they are used to present works and ideas, to discuss new products, to determine trends, to socialize, and to initiate co-operation and collaboration. Conferences pay special attention to the assembling and the provision of scientific contributions.

Publications document the scientific contributions of the conference. Publications take various forms with individual strengths and weaknesses in distribution, storage, capacity and access capabilities. The *abstract book* documents the scientific contributions of the conference. Traditionally printed, abstract books nowadays are distributed as net publications. The *conference web site* allows for interactive structured access to the abstracts along date and time, type of presentation, topic and presenter, embedded in the overall scientific program of the conference.

Data collections incurred and managed in the course of one conference and/or consecutive editions of one conference features heterogeneous material with metadata and multitude of relationships. Entities in a collection comprise amongst others, abstracts, papers, posters, presentations, authors, sessions, and topics. Data and document collections comprise two general types: self-contained documents that can be considered complete and well-established (for example, presentation slides, posters, the printed conference abstract book), and the multitude of data, texts, images, and document parts gathered or produced in the course of the conference. This material includes amongst others, organizational data including conference participants, presenters, events, sessions, talks, and topics, as well as structured text information from conference contributions, especially abstracts and their tables and figures.

For a conference scientific contributions are accepted, indexed and reviewed. Speakers get invited, a scientific program is set-up, categorized and linked thematically.

2. Information Life Cycle Model

Context data of digital objects evolve in different phases of existence. Context is guided by the processes in which the digital object is created, preserved, accessed and reused. Today, archives often depend on deriving metadata from the digital object obtained from the producer together with a minimum metadata set called-in by the archive. A good share of the imprint of the digital object gets lost during its transit into the archive. Opening-up the context of digital objects requires the capturing of context during all life phases of the digital object. Those life phases of a digital object are modeled in the archive-centric Information Life Cycle Model, depicted in Figure 1. The model distinguishes five relevant phases:

- **Creation:** new information comes into existence.
- **Assembly:** denotes the appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the Designated Community. Assembly requires in-depth knowledge about the Designated Community in order to determine objects relevant for long-term preservation together with information about the object required for identification and reuse some time later in the future.
- **Archival:** addresses the life-time of the object inside the archive.
- **Adoption:** encompasses all processes by which accessed archival packages are unpacked, examined, adapted, transformed, integrated and displayed to be usable and understandable for the consumer. This includes also emulation activities if needed. The adoption phase might be regarded as a mediation phase, comprising transformations, aggregations, contextualisations, and other processes required for re-purposing data.
- **Reuse:** means the exploitation of information by the consumer. In particular, reuse may be for purposes other than those for which the Digital Object was originally created. Reuse of Digital Objects can lead to the Creation of other, novel Digital Objects. Reuse also may instigate the addition or updating of metadata about the Digital Object held in the archive. For example, annotation changes informational content and affects the relationships existing between the Object and other Digital Objects.

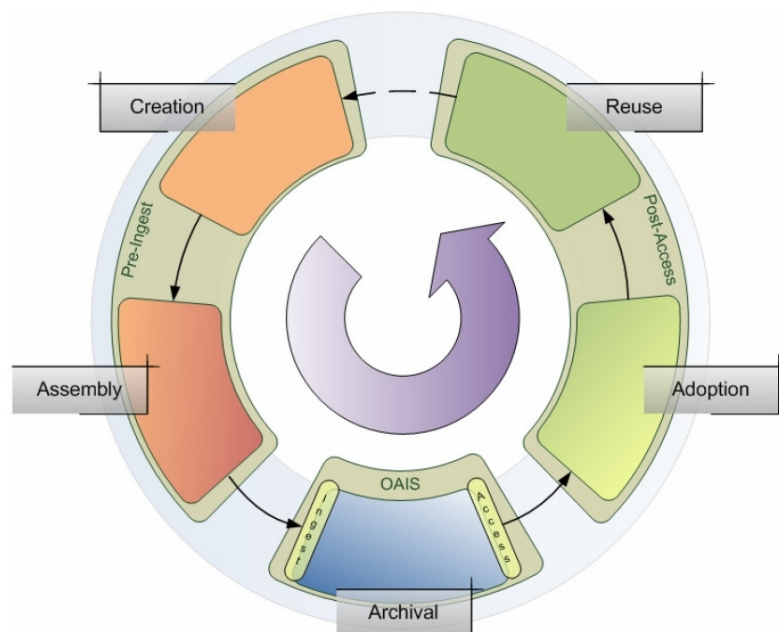


Figure 1: Information Life Cycle Phases

3. Context and its Representation

The pursued approach regarding *context* is Digital Preservation (DP) centric, following the Open Archival Information Systems (OAIS) Reference Model [1]. OAIS is a framework of terms and concepts providing a standardization of archival systems.

Context accords to the *interrelated conditions in which something exists or occurs* [3]. This expresses generally what all context definitions have in common. This statement implies for digital resource management, that the context of a digital object is complex, possibly

containing concepts which are shared with other objects. This might be the process environment in which they are created, the associated actors, resources and information objects and also the preservation environment in which they are stored.

Furthermore, different domains and different scenarios have different requirements towards a context definition. Currently six content components are distinct in the context approach of the SHAMAN project: *Document Context*, *Production and Reuse Context*, *Preservation System Context*, *Modeling Change Context*, *Social- and Enactment Context*.

The most important context component for scientific publishing is the *Production and Reuse Context* (PRC). This context component corresponds to the producer and (anticipated) consumer environment, i.e. the respective designated communities creating and accessing digital objects. The creation environment includes the actors and resources involved, but also a formal representation of the organizational and technical processes carried out in the production of a digital object. To re-trace information paths, the representation of the production context has to be maintained during the transition from the production into the preservation environment. The reuse of preserved digital objects depends on a proper description of the significant properties and the associated domain-specific knowledge. This description allows for the efficient access and usage even from outside the designated community.

Especially in the PRC it is obvious that context is not only defined by the digital objects themselves, but it is also defined by the processes, in which they were created, preserved, accessed and reused. Domain-specific groundings provide interfaces to the relevant concepts and topics of the designated communities addressed, in addition to formalizations of the organizational structures involved, including associated role assignments. Concluding from this three distinct concepts are encountered, which are strongly involved in defining context. Those are:

- **Domain:** the concepts specific to the domain and their relations. For instance in the domain of scientific publishing: *Abstract*, *Abstract Book*, *Presentation* or *Supplement*.
- **Enterprise:** the structural layout of an organizational environment. For instance in the domain of scientific publishing: *Affiliation*, *Persons* or *Roles*.
- **Process:** the processes and their associated activities, including information about their implementations (service invocations): *Submission*, *Indexing* or *Reviewing*.

If context data should be preserved over time, a model for representation and organization of data is required. As a structured representation form of *concepts* and their *relations*, the usage of ontology is appropriate. An ontology represents *concepts* and their *relations* to one another. This could be seen as a formal model of a specific domain (see e.g. [5]). Ontologies are used to establish a common understanding about knowledge existing within a domain. One important aspect of ontologies is that they formally express the semantics of each element contained, enabling individuals and machines alike to access and process the knowledge represented. Rules and inference (or reasoning) mechanisms can be employed to derive new insights, i.e. making so far implicitly existing knowledge explicit.

The ontology used in SHAMAN is conceptually structured in the three sections Domain Ontology, Enterprise Ontology and Process Ontology. Those ontologies are consolidated through the ABC ontology [8], which was formally developed to model resources and their spatial, temporal, structural and semantic relationships.

4. Context-oriented Information Retrieval

Basing on the context notion and the representation of context as described in the previous sections, retrieval could be extended in two ways: firstly through the creation of an additional

full-text index, containing the indexed context data and secondly through retrieval mechanisms on base of the relations between archived objects. Such relations between archived objects evolve through similar context attributes values. Those attribute values in the domain of scientific publishing are for instance the same author, the same conference, the same reviewer or common keywords. These data should be accessible through query, browsing with visualization support. The result of such a context oriented query is then not restricted on the archived objects; rather this could be a set of context data.

Context data in the domain of scientific publishing which can be expected to aid the retrieval of relevant publications for the purpose of scientific reuse are, for example

- Representation types such as abstract, presentation slides, poster or full paper;
- Embedding in the world of scientific discourse along citation nets, roles, interest and competence profiles of persons and organizations, and discussion threads;
- Implicit and explicit relationships to other documents like review reports and conference reports.

Those data could support, for instance, the retrieval of information for a state of the art research. Once a first relevant publication was found it could be used as the starting point to search for similar publications. Similarities according to publication context are, for example, but not limited to: publication origination from conferences with similar subject focus, by origination from the same conference, its conference sessions, its tutorials or its keynotes. Furthermore, it could be valuable to find publications with the same key words, publications which are referred to the source publication or the publications that refer to the source publication.

Different approaches for defining the concept *context* exist in IR. A user centric approach has been done for instance by Järvelin et al. in [6]. They stated that context is given through dependencies in time, place, history of interaction, task at hand and some other factors. Another approach towards a context definition in IR has been outlined by Cool et al. in [2]. They classify IR context in four different levels, namely: *information environment*, *information seeking*, *IR interaction* and the *query* level.

Some conceptual and implementation work on context based IR is already done. Melucci for instance presents in [9] a context model and the application of the model for ranking. Some context based IR support tools are implemented in Daffodil. This is an experimental system for IR and collaborative services in the field of higher education for the domain of computer science and others [4]. Daffodil comprises, for instance, an Author Net, which depicts relations among authors stored in a database and is used for ranking and the search for central actors in a set of documents or central actors for a specific author. Daffodil furthermore implements a Citation and Co-Author Browser, which are similar the Author Net, as well as an adaptive suggestion tool, which is based on the current situational user context [7].

But even if some particular solutions towards context oriented retrieval are implemented yet, the retrieval in preservation systems access lacks of offering a holistic model of digital object context for different domains and the preparation of context data for usage in retrieval.

For such a context oriented IR process it is essential to:

- define a holistic and adaptive context model, in order to serve the requirements of different domains
- provide mechanisms to capture relevant context data during the ingest phase
- prepare the context data in order to make them usable for retrieval
- offer an appropriate query- or browsing format in order to query the context data
- offer an appropriate way for presentation

The support of all those requirements is a task for future scientific work.

Conclusion

In this paper the advantage towards context oriented IR in an archive information life cycle is motivated. A context notion on basis of ontology is presented in order to model the context of preserved digital content. The ontology based representation provides valuable additional information for IR through the description of relations. By means of the archive-centric information life cycle model, the important phases for capturing context are presented. The domain of *scientific publishing* was used to illustrate the usage of this retrieval approach.

References

- [1] CCSDS. Reference Model for an Open Archival Information System (OAIS). Blue Book 1, Consultative Committee for Space Data Systems, January 2002. Recommendation for Space Data Systems Standards, adopted as ISO 14721:2003.
- [2] Colleen Cool and Amanda Spink. Issues of context in information retrieval (IR): an introduction to the special issue. *Information Processing Management*, 38(5):605–611, 2002.
- [3] Merriam-Webster Online Dictionary. context; cited 30.04.2009. "ONLINE" <http://www.merriam-webster.com/dictionary/context>.
- [4] Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. Daffodil: An Integrated Desktop for Supporting High-Level Search Activities in Federated Digital Libraries. In *Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002*, pages 597–612. Springer, 2002.
- [5] Nicola Guarino. Formal ontology and information systems. In Nicola Guarino, editor, *Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98*, pages 3–15, 1998.
- [6] P. Järvelin, K. & Ingwersen. Information seeking research needs extension towards tasks and technology. "ONLINE" <http://InformationR.net/ir/10-1/paper212.html>, 2004.
- [7] Claus-Peter Klas, Sascha Kriewel, and Matthias Hemmje. An Experimental System for Adaptive Services in Information Retrieval. In *Proceedings of the 2nd International Workshop on Adaptive Information Retrieval (AIR 2008)*, October 2008.
- [8] Carl Lagoze and Jane Hunter. The ABC Ontology and Model. In *Dublin Core Conference*, pages 160–176, 2001.
- [9] Massimo Melucci. A basis for information retrieval in context. *ACM Trans. Inf. Syst.*, 26(3):1–41, June 2008.
- [10] Ute Schwens and Hans Liegmann. Langzeitarchivierung digitaler Ressourcen. In Rainer Kuhlen, Thomas Seeger, and Dietmar Strauch, editors, *Handbuch zur Einführung in die Informationswissenschaft und -praxis*, volume 1 of *Grundlagen der praktischen Information und Dokumentation*, chapter D9, pages 567 – 570. München : Saur, 5., völlig neu gefasste Ausgabe. edition, 2004.