On the Relation Between Qualitative Digitization and Library Institutional Identity

Abstract
The paper highlights and discusses concepts and practices of national library digitization. Two conceptual models are suggested in order to help strengthen scholarly analyses of digitization practices within libraries: the distinction between quantitative and qualitative digitization, on the one hand, and the prism metaphor for understanding the nature of qualitative digitization, on the other. Qualitative digitization, understood as a document representational practice, is defined as a knowledge organization practice. These concepts and models are then put in relation to the issue of national libraries and institutional identity. By combining research within the fields of KO, digitization and institutional identity, the paper points to a way of addressing empirical issues in all three fields of research. Special focus is on the very definition of qualitative digitization as a KO practice and the practice of selecting and digitizing documents suitable for the development of institutional identity within libraries.

Aim
The aim of this paper is to discuss and rethink concepts of digitization and knowledge organization (KO) practices in relation to cultural heritage digitization and library identity. By referring to practices in Scandinavian national libraries and institutional identity theory, claims are made that document digitization and markup should be considered as KO practices and that they play an important part in shaping cultural heritage and the very identity of national libraries as institutions. Rather than producing answers, our aim is to provide a basis for further questions and thus create a platform for future research. It is our belief that placing KO practices in new contexts and in relation to current trends in the ALM (Archives, Libraries, Museums) sector, we may contribute to a relevant scholarly discussion in Library and Information Science (LIS).

The paper starts off by discussing different forms of digitization as practices of KO. Turning to research on institutional identity relevant for the ALM sector, conclusive suggestions are made on how to develop research on the relation between qualitative digitization and library identity as part of the theoretical development of KO.

Digitization as Knowledge Organization practice
Let us begin by making two clarifications. First, we make a distinction between quantitative and qualitative digitization. The former, sometimes referred to as “mass digitization” and exemplified by the Google Books, the Open Content Alliance or the European i2010 efforts, aims to digitize massive amounts of documents (thus an all-inclusive strategy) using automated means (Coyle 2006). For pragmatic reasons, mass digitization has to disambiguate the transmission phases, minimize interpretation and flatten out the digitization process into a two-dimensional linear affair. Qualitative (or critical) digitization (henceforth shortened as QD) on the other hand involves deep text encoding, critical image and text editing, and rich information assignment to critically selected documents, a more exclusive strategy. Secondly, when we are referring below to text encoding, we are talking about descriptive strategies using element-rich XML-compliant schemes such as the TEI (Text Encoding Initiative), where the contents and structures of a document are assigned descriptive labels, as opposed to procedural and presentational strategies, where
encoding is used primarily for visual effects (Renear 2004). The TEI scheme is appropriately employed within qualitative digitization projects, while quantitative digitization uses simpler text encoding, if at all, merely for presentational (and in this context more trivial) purposes.

Cultural heritage digitization in the ALM sector has relevance for, and is being studied within, several areas within LIS (e.g. collection development, digital libraries, book history, open access and publishing studies). Knowledge organization (KO) is arguably the LIS area where digitization, and in particular QD, proves to be most fruitful for testing and development. For instance, digitized document collections and the way they are internally organized offer considerable challenges to fundamental bibliographic concepts and models (Dahlström 2006, ch. 2; Renear et al. 2003). Further, digitization practices prove to be intimately engaged in KO practices — to the extent that we argue for considering QD as partly a KO practice. A major area of KO can be called document representation (DR), exemplified by metadata production, bibliographic practices and descriptive cataloguing.1 In DR, documents and their components are analyzed, labels and other metatexts are assigned to them, and textual parts of documents are transmitted into a new, target document, a representation. This representation can range from being quite brief, as a catalogue post, to being a more or less full simulation of the departure document, where specific codes are assigned to the document sections at various granularity levels (Dahlström 2004, 21). QD using detailed descriptive encoding is another instance of such DR.

QD is connected to cataloguing and metadata in three ways. Firstly, proactive as well as reactive digitization (Lee 1999) often necessitates a new and separate cataloguing of materials that have so far not been catalogued at all, usually because the materials are component parts of other documents, e.g. single sheets or graphic objects. Secondly, sophisticated descriptive text encoding demands a relatively exhaustive metadata account of both the original and the digital document (in TEI this is performed using a so-called TEI Header). Thirdly, the very application of a markup scheme and its elements means structuring, labelling and transmitting (parts) of documents in a way much similar to other DR. Such DR practices within QD are conditioned by norms and strategies that are either described in policy principles (“best practices” and guides) or formalized in standardized schemes.

**Qualitative Digitization as prism — a tentative metaphor**

Where quantitative digitization implies a simple linear transmission model (simply having the book scanned and putting the resulting image files on the web), qualitative digitization involves intellectual, interpretative and multi-faceted measures at several phases of the digitization chain. Such critical, interpretative measures have for long been recognized by textual critics and scholarly editors as the basis of their practices, and QD, applied rigorously, indeed becomes an application of scholarly, critical editing. Perhaps an even more suitable term would therefore be critical digitization.

The input phase of QD includes e.g. policy making/adaptation and the identification, selection, collation, analysis, description, cataloguing and classification of source documents. The preparation phase involves such tasks as image and text capture, transcription/OCR, conversion, image enhancement, intellectual text encoding, database construction, annotations, as well as metadata work with the digital material. The output phase,

---

1. With subject representation and IR as the two other major KO sections.
finally, has to do with activities such as manual or automatic transformation (e.g. using XSLT), versioning, layering, linking, publication, availability, archiving, optimization, added values, feedback, marketing, evaluation, and the important questions of use and re-use. What comes out in the end is not a simple clone of the analogue source, but an enriched, interpreted, changed and new object that to a certain degree reflects but also comments on the source. A user approaching a digitized collection thus faces an interpreted and argumentative material that answers the particular questions the digitizing agent has asked during the digitization process. As noted by Rockwell (2003, 215) however, researching a digitized material means being able to subject it to new and unanticipated questions — a scholarly ideal that is conditioned and, arguably, hampered with each intervening interpretation.

In this sense, QD can be understood as a *prism*. Prisms reflect, deflect, or transmit light, and are used in e.g. cameras to split light into a spectrum of colours and with the aid of mirrors create a representation of an object by producing a simulation of it. Similarly, QD splits a more or less cohesive analogue object into a spectrum of files that vary in category, function, extension and level. In QD, one would e.g. end up with packets of files such as raw texts with stand-off markup, or text files with embedded marked up (.xml or .tei), document type definitions — i.e. markup rules for a genre or document type (.dtd), style sheets (.css), transformation instructions (.xsl), facsimile source files (.tiff), image delivery files (.gif or .jpg), metadata (in RDF or teiheaders), databases etc. These different spectrum files may then be edited separately, combined and re-combined at will, each constellation bringing a new representation of the source as an altered simulation.

The agent performing QD, such as a library, faces most of these possibilities at various phases in the process, and has to make choices of what to select and what to leave behind, what to enrich intellectually, and what to make available to which audience in what form for what purposes. The library’s implementation of DR/KO practices such as versioning, bibliographic relations, text encoding, document organization and distribution is instrumental in that process, and conditions the possible ways the material selected can be made available, such as in what formats and various outputs, and what possible uses and re-uses can be made of the material. In so doing, the digitization design legitimises particular kinds and levels of materials at the expense of others and addresses particular needs and audiences at the expense of others — a question of symbolic power.

The library’s decisions and overall QD strategy for selection, preparation as well as availability has significant bearings on the overall shape of the accumulated digitized library collections. Qualitative is more likely than quantitative digitization to make the digitizing library attentive to the intellectual, interpretative features of cultural heritage digitization. Intellectual DR practices such as text encoding and other forms of descriptive metatext labelling are important in shaping the prism trait of QD, the emerging pattern of the national digital library holdings, and to some extent therefore the cultural identity of a national library.

**National libraries and institutional identity**

Issues on institutional practice and identity have been dealt with in several disciplines, most prominently perhaps in political science and sociology (March & Olsen 1989, Peters 1999). Within LIS it has been brought up by Frohmann (2004), and identities of libraries have been discussed by Audunson (1999), Goodrum (1980), and Hansson (2006). Most of these studies take their departure in the relation between norms, rules and associated
practices in defined types of libraries such as public libraries or national libraries. The identity of an institution manifests itself through daily practices and decided policies. The definition of institutional identity however goes beyond this. A national library forms its identity in relation to the idea of the national state which it is supposed to reflect in terms of collection development, for instance through legal deposit regulation, but also through the relation it has to the rest of the library sector. National identity is mirrored in the identity of institutions. In Europe there has, by tradition, been a sharp division between e.g. public libraries and the national library, while in many younger national states, such as many of those in Africa, we can see a much closer integration of the different parts of the library sector (Hansson and Kawalya 2007).

KO practices such as document representation and subject analysis are significant tools in library practice that help define this idea, the set of values underlying the institution (Andersen 2004; Hansson 1999). Kjellman (2006) shows how institutional practices directly influence the practice of KO in her domain analytical study of the picture collection of the National Library of Sweden. Reversely, she concludes, the practice of KO constitutes a major influence on the general institutional practice of the National Library.

Over time we can see different practices emerge and perhaps the most prominent one in later years is that of digitization of various parts of the collections or specific documents. This practice contributes significantly to the totality of identity-shaping practices and, as it is relatively new, to the development of the identity of library institutions in contemporary society. The shaping of an identity is on the one hand based on practices that come out of the formulated mission of the library, but it is also a result of intentional decision-making regarding how the library wants (or expects) to be perceived by the outside world. Depending on the points of departure in individual cases, there are thus many ways of letting the process of digitizing collections contribute to the shaping of a developed identity of the institution. Two examples of the variation of decision-making and digitization practices are the national libraries of Norway and Sweden.

In Norway, the National Library is pursuing the practice of digitizing, more or less, their full collection. The library today has over one million documents digitized. This practice is based upon quantitative digitization, which, as suggested earlier, hinges to a lesser degree on the intellectual processes of KO than does qualitative digitization.2 The National Library of Sweden, on the other hand, is presently engaged in a high degree of selection and refinement in the individual expressions of the digitized document, produced in small-scale projects. This obviously means applying qualitative digitization, where unique and historically high profile documents are chosen from the collections and subjected to detailed text encoding and analysis.3 In all aspects, the digitizing practice of the National Library of Sweden follows the intellectual practice of interpretation that is so characteristic of qualitative document representation. As such it is an obvious renewal of traditional KO practices. Why is this so important to stress? It is so because of the impact KO has on the identity of libraries as social institutions.

---

2. It should be noted that the National Library of Norway also works with more limited, qualitative and critical digitizing projects in parallel to its mass digitizing policy. An example of this is the collected manuscripts and letters of Henrik Ibsen, http://www.ibsen.uio.no.
3. Two examples of such qualitative digitization projects of the National Library of Sweden concern the Codex Gigas (http://www.kb.se/codex-gigas/eng/) and the book Suecia antiqua et hodierna (http://www.kb.se/samlingarna/digitala/suecia/) — the latter being, as a case in point, a classic and monumental expression of the formation of Swedish national identity during the 17th century.
From this point of view, it is clear that these two approaches both contribute, but in very different ways. Digitizing practices contain more or less conscious processes of interpretation. When ALM institutions engage in mass digitizing, they are likely to make use of a simplified media model which trivializes cohesive documents and their transmission, treating the latter as mere cloning. Kjellman (2006, 239 f.) further observes that whereas museums tend to consciously discriminate and select, the "totality ambition" of libraries when collecting materials results in notions of neutrality, which tend to hide the institution’s discriminating mechanism. Such a tendency is obviously fuelled by quantitative digitization. Qualitative digitization on the other hand is more likely to emphasize the interpretative aspects of the KO principles underlying the process, as described by the prism metaphor. The choices made are conscious, and the result usually defines not only the actual process in a narrow technical sense, but in an institutional and socially relevant way as well.

Cultural heritage (CH) digitization in libraries has two main objectives: preserving the materials and making them available to current and future audiences, and thus also to support use and re-use of the materials within education and research. To support such objectives, the issues of what material is being selected as input and what kind of material is being made accessible as output — and how — become crucial factors. The selected documents fill a symbolic function as representatives for the contents of the whole library. So does the way in which they are bibliographically organized. The selection of documents to be exhibited in digitized forms is crucial in forming the identity of the institution as accessed through the web. KO practices are among those that form the basis of this identity-shaping selection of cohesive documents to be fragmentized in the elaborate digitizing process. The way in which this is done in practical library development, however, needs to be further analysed.

Conclusion
The relation between digitization, KO practice and the development of institutional identity within ALM institutions has not been subjected to analysis in any major way. At the same time, digitizing projects, both qualitative and quantitative, are put forward as crucial to the libraries, archives and museums involved. Digitization is seen as a way for these institutions to keep up with the Zeitgeist and often provides a major argument for the relevance of them in contemporary society, both on a local and regional level. In our coming research efforts in this field, we will focus on the selection and accessibility phases. More precisely, we will address questions such as:

• How do the publication formats include or exclude audiences, and how do the KO methods applied support use and re-use of the digitized material?

Relevant problems to study include text or image capture strategies; granularity and versioning; text encoding and transformation; restricted versus “open source” access; scalability, modularity and longevity (cf. Besser 2004; Bodard and García 2008; Dahlström 2008; and Seaman 2003). Those studies will require analyses of what kinds of material have been selected and why, at what audiences they are targeted, and what possible implications that might suggest in relation to the identity discussion implied above. In other words, when focusing on national libraries and on the issues of selection and availability,

---

4. Individual examples of such analyses can of course be found. See for instance Dalbello 2004.
ity as manifest practices in relation to an identity which conjuncts with the national state, several issues of concern emerge:

- In what way do the digitized and organized material manifest issues of identity in terms of functionality and availability for various kinds of users?

The issues raised in this paper should be seen as part of a discussion of the impact of new practices in librarianship. Practices based on KO principles are the very core of this development. Using the distinction between quantitative and qualitative digitization along with the prism metaphor for combining relevant discussions of institutional identity, practical digitization and KO principles, provides us with an opportunity to raise questions concerning these socially and institutionally highly relevant relations in a more informed way.

Acknowledgement
We would like to express our gratitude to the two anonymous reviewers of this paper. Their comments helped to improve its content.

References


