Deschooling information: exploring uses of digital archives in research and higher education

A proposal for a research project by:
Monica Langerth Zetterman
Ph D candidate
Dept. of Education, Uppsala University
2003-05-06

Contents

Introduction

In my Ph D project I will explore how teachers and students could be given opportunities to separate what is learnt from how it is taught in higher education. Given the present and emerging possibilities, in tools and methods aimed to separate the structure and the organisation from the actual content, my overall aim is to study how we can make information applicable for different purposes and uses in research and educational practices. In trying to reach that aim I will explore - through case studies and content analysis – actual uses of tools and methods for electronic production, preparation[1] and uses of text-based content and material[2]which constitutes or are parts of digital archives. The focus for my exploration of digital archives and practices will be research communities in the humanities but to some extent also in the natural sciences.[3] The aim and the questions but also the methods and the procedures I propose in this PM are preliminary suggestions and an invitation to a discussion of my initial ideas – some of them more elaborated than others.

During the last decades information technology has changed the research practices of many scholars. One important aspect is the augmented opportunities for co–operation and exchange within research communities where the Web has transformed the way in which scholarly knowledge might be produced and disseminated. The transformation has been most notably in the sciences and humanities since the adoption of digital technologies for research, analysis, communication, and teaching, has been more slowly in the social sciences, even though there has been progress in these fields. [4] This development is of immense interest and importance for anyone concerned about how to acquire, preserve, and make accessible content coming from different research practices.

Many communities, e.g. within universities, produce, store and share very diverse content, spanning from syllabi and lecture notes to articles, papers, simulations or research results stored in databases, which is of use to the community members. Communities are to be taken as communities of practice, according to the categorization of used by e.g. Lave & Wenger but also similar to the notion of "epistemic cultures" as utilized by Knorr Cetina.[5] The digital archives could be valuable learning resources for old and new members of the

[1] There are several aspects involved in and related to the preparation and production of digital content and my use of the terms will include several important ones. The aspects I include in "preparation and production" concerns both research and education and will be elaborated more or less in my studies; issues of access, availability and retrieval, issues of possibilities to manipulate the information and issues of how to organize and structure the content. There are also issues I am leaving aside and here I am thinking of issues relating to cultural heritage such as responsibilities and standards for (long term) preservation and storage or metadata for cataloguing and classification which is an important concern not only for archives and libraries, but to all of us who are using these services. For a comprehensive overview of these issues, and many more, see Lazinger, 2001. Nevertheless, I have no intention of cover all real or possible issues at hand in any comprehensive or full scale basis in my forthcoming studies – that would be too large and to difficult for a singe Ph D project. However I hope that this presentation (PM) will illuminate my choice of aspects and concerns I am in fact going to explore.

[2] Foremost electronic texts from different kinds of scholarly contemporary (re)sources and/or historical documents such as, books, articles, reports and to a lesser extent other kinds of data (annotations, database links, numerical sequences).

[3] The communities I will study are the humanities computing community and especially the TEI community and bioinformatics in the natural sciences.

[4] For comprehensive overviews of pertinent implementations and initiatives of electronic text and digital archives in the humanities see e.g. Condron, Fraser & Sutherland, 2001, Hockey, 2000, and Sutherland, 1997 and see e.g. Lazinger, 2001 for an overview also including the social sicences.

[5] Epistemic cultures are those practices of arrangements and mechanisms which, within a given field, make up how we know what we know. Cf. Knorr Cetina, 1999.

communities. However, the rapidly growing information on the Web and intranets causes problems when searching for and accessing information. Likewise, the same problem occurs when individuals are trying to organize and share resources[6]. One problem with the growing digital archives is that content often is moulded into proprietary and application dependant formats and therefore difficult to retrieve, access and reuse in any other context than that in the one in which they are produced.

Ten years ago, individual researchers, research groups and archivists typically used proprietary non - portable formats for their depositories of digitalized sources, for example in the humanities electronic versions of printed texts, transcribed manuscripts or language corpora. Today there are de facto encoding[7] standards, such as the TEI DTD (the Text Encoding Initiative Document Type Definition)[8] within the humanities, which permits scholars to create, use and share collections of well structured high quality digital sources. The same kind of development has taken place within most scholarly domains, as well as in many industries but not in the same extent in the teaching practices in higher education. This means that the educational milieus are lagging behind the research communities. However, most teachers and students do not take part in this development. They are still referred to printed material, poorly structured HTML pages or non–reusable PDF files. The qualified digital sources they encounter are more often than not only available on proprietary systems and formats that may function well as long as the material stays on the web site or on the CD-ROM. Students and teachers meet with problems as soon as they try to incorporate material into their learning environment, for example in order to accomplish projects or to create tailor–built courseware. Course content is, thus, not sufficiently integrated into the learning environment.[9]

Why then, deschooling information? When Ivan Illich launched his campaign 30 years ago against the schooling of society, the educational system was the first in a row of several attacks (titled accordingly) on the augmented professionalism and the top–down management of schools making students powerless and therefore prevents people from learning. Illich proposed the need to replace schools with libraries as the natural arena for learning to expose the authoritarian hidden curriculum of education.[10] The idea that schools could be replaced by *networks of information* proposed by Illich I think is relevant in some respects especially in terms of what kind of opportunities it might bring for educational practices. Firstly because the flows of information have become easier to reach e.g. by uses of information technology and secondly because greater possibilities for both producers and users to manipulate and use the content of information. Indeed, that may be part what Ivan Illich was looking for when he wrote in *Deschooling Society* and as part of his proposal of learning webs, Illich thought of ways to "[...] provide the learner with new

[6] What is a resource? It might consist of almost any kind of representation and a digital resource is any object or location that has a *unique identifier* and can be digitally stored, accessed and distributed via a global/local area network. A resource could be e.g. web pages, course modules, lecture notes, reports, papers, digital archives, a record in database or a student digital portfolio.

[7] The purpose of encoding a text or other kinds of data is to provide information in order to add some kind of functionality such as assisting a computer programs to perform actions on that text. A common kind of simple encoding is HTML which allows the web browser to display a given resource such as a text without showing the actual encoding i.e. the tags.

[8] The guidelines from the consortium TEI (Text Encoding Initiative) are the dominating de facto-standard for encoding textual resources such as digital critical editions of literary and historical sources. See http://www.tei-c.org/

[9] These issues have been presented and discussed by Broady in "Det nya handbiblioteket" (1995), "Content design. Methods and tools for the creation of portable hypermedia archives" (1997) and "Digitala arkiv och portföljer" (2001) but also to some extent by other, for example Julia Flanders in her article "Learning, Reading, and the Problem of Scale: Using *Woman Writers Online.*" (2002).

[10] See Illich in "Deschooling Society" (1971a) but also Illich, 1970 and 1971b.

links to the world instead of funnelling all educational programs through the teacher."[11] He envisaged all sorts of means in the learning web, primarily technological ones, to access information autonomously, to communicate with others, exchange skills and to make the learners own ideas available to others who might make use of them and so on and then he coined the term: *learning web*, as a term for the liberation technology he foresaw.

Hence, through my title which is inspired by Illich I want to emphasize what this study is about: open kinds of experimental practices and principles of content design and the accompanying tools and methods. Meaning for instance freedom of use or freedom to manipulate the software or the content but also sharing content and working together for improvement or problem solving. Within the software domain, the concept incorporates the notion, open source, which basically means that the source code is available or open for users and other programmers to read, to use and possibly reuse in different projects.

Another prominent example of an open and experimental practice is the Oxford Text Archive, OTA, where they archive texts for future generations of scholars.[12] The OTA originally founded by Lou Burnard in 1976 to stop people from duplicating the work of other researchers and instead collect copies of this kind of material to share it with anyone in the world who needed it. [13] At the OTA they encourage producers and contributors to make texts as freely available as possible for (private/individual) research and teaching, but not for any commercial business although this is up to the original depositor to decide. OTA started long before the web and they currently have more than 2500 encoded full text high quality titles in their archive. Many of the texts collected at OTA are encoded according to the TEI guidelines mentioned above, which is my third and last example of openness.

The TEI community's original overall aim was to develop interchange guidelines allowing projects to share data and theories about data and to promote the development of common tools. More than 100 persons, from many different specialities and disciplines have been working hard since 1987, when the TEI project was launched, to produce and rewrite the guidelines.[14] This effort have been very successful and the TEI guidelines are now widely accepted as the standard interchange format for textual data and there are quite a few full text archives and projects currently using the TEI guidelines.[15]

Aims of the study

The overall aim with my study is to explore how we can make information applicable for different purposes and uses in research and educational practices. And the subject will be to strive for a systematic exploration of digital content design[16] in research communities with the auxiliary practical goal that such understanding might be valuable for educational practices in higher education in terms of how teachers and students could be given opportunities to separate what is learnt from how it is taught in higher education. My primary focus is on uses of tools and methods for preparation and uses of electronic text-based resources, since text is the dominant form of communicating knowledge in research and education. I will also, to a far lesser extent, study the preparation and usage of other

[11] ibid. chapter 7.
[12] See the OTA website at http://ota.ahds.ac.uk/
[13] A researcher can spend five years typing in and encoding a text in ancient Greek, and if this is not known, another person might start doing the same thing.
[14] TEI's *Guidelines for Electronic Text Encoding and Interchange* (TEI P3), first published in April 1994 (more than 1000 A4 pages) and have been revised a number of times since 1994.
[15] See http://www.tei-c.org/Applications/index.html
[16] When using the notion of content design I refer to: digitized preparations by encoding of electronic texts, and applications of different kinds of guidelines, tools and methods.

kinds of data, such as numerical sequences and database links since I am planning to do a comparative case study of two different research communities, the Text Encoding Initiative in the humanities and the bioinformatics in the natural sciences.

Research questions

The interrelated but yet standalone studies will be guided by the following questions and sub-questions:

- How might teachers and students benefit from emerging mark-up and metadata recommendations and standards?
  How might digital archives be designed allowing teachers and students flexible modes of access to and use of the content therein?
- How does the terrain look like of in the uses of digital archives for humanities and social science research and education?
  How is the complexity of designing and preparing different kinds of content handled? What strategies, principles and interests inform their orientations?
- How do research communities make use of digital archives and web repositories?
  How do researchers actually work with their tools and methods in the humanities communities on the one hand and the natural sciences on the other hand?

The problems, context and delimitations

Above I have tried to argue for why it might be important to study how digital archives and digital content are designed, managed and utilized in research but not yet discussed in what way it might relate to teaching in higher education. Below, I will discuss some pedagogical aspects on the relation between personal portfolios and shared repositories and how digital archives might be designed to allow teachers and students to access, retrieve and use that content in different ways. I also want to give a brief overview of emerging mark-up and standardization efforts and initiatives with relevance for educational settings. And last but not least the issue of use of digital archives and the practices of content design in some research communities. This could of course be studied in a number of ways and I will try to unfold the issues inherent in the questions above by using three interlinking strands. My intention is that these strands will serve as means to discern what kind of assumptions, levels and relationships I could be dealing with in my forthcoming studies.

Content: the nature of information utilised and studied by scholars and teachers can be almost anything and the information can be studied for many *different reasons and purposes*. Thus, digitization of content, information and source material involves some fundamental decisions: what is considering *being the data*.[17] In this case, content will primarily be equivalent with texts but to some extent other kinds of data as mentioned above.[18]

[17] When Willard McCarty (2002) discuss what happens when source materials (in the humanities) are computerized he stresses the *interaction* taking place during the process of selection and choosing what is consider to be the data. The difficult part in construction of computerized source material is the inevitable *difference* between human knowledge of and declarative statements about specific objects (i.e. the data).

[18] I will not – at this stage anyway – discuss or attempt to elucidate what a *text* is or might be. I leave this discussion for the moment and instead treat the notion of text from a pragmatic and rather unproblematic view since my interest is within uses, manipulation and applications of electronic texts. Namely, when talking about texts I refer to signs organised as information and representation of any content. A linguistic definition could be: "a text is a sequence of paragraphs that represents an extended unit of speech" (Glossary of linguistic terms, http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/Index.htm [2003-04-07]).

Tools and methods: what kind of tools and what kind of methods and their accompanying underlying principles are guiding how to *prepare*, *organise*, *retrieve*, and possibilities to *manipulate* the content. One very important distinction here is the difference between *internal* and *external* use of metadata.[19] External metadata might consist of information added outside the actual recourse in order to be able to e.g. classify, order, organize, store, rearrange or facilitate resource discovery and an example of this is different library catalogues and databases.[20] Internal metadata on the other hand is to provide information which is embedded in the text - variously called either encoding, mark-up or tagging - which will assist computer programs to perform functions or manipulations on that text.[21]

Experimental practices:[22] the communities in which the digitally manipulated information are produced and used provide the third aspect for the exploration on production and uses of digital archives in research and higher education. The analogy of "experimental practices" to the experimental sciences might be useful to this project since I intend to study the actual work in some research communities characterized at large by experimentation as such.

On the relation between personal portfolios and shared repositories - educational uses of digital archives

Many systems and platforms for computer support in higher education comprise fairly sophisticated tools for the communication between teachers and students and between students, for management of the course and the assessment of results etc.
While the content – the courseware, the reference material, the teachers' contributions or the students' papers – is often poorly structured and lack appropriate encoding or metadata. Besides, much research on uses of ICT in higher education focuses such top-down matters mentioned above or teacher/student interaction, teaching methods or learning outcomes.[23] Some studies, slightly related to this project, focus on how digital libraries are used.[24] Some critics, such as Laurillard stress that these R&D efforts seldom scale down to the teaching and learning practices or up to institutional level.[25] Besides, the content issues are neglected and less numerous are studies investigating and exploring how technology might

[19] A common definition of metadata is: "data about data". This definition is too terse and vague to take us very far, indicating a diffuse environment of use. Thus, it might be useful to extend and be more precise in a definition to its intended use or environment. This is not my task here though, and therefore I will try to be specific in which way I use the term "metadata". A simple example: a search for "definition AND metadata" in Google returns more than 200 000 hits from various sources and different intended uses such as technical, educational or archival definitions. Many resources, such as documents, holds self-descriptive metadata such as title, author and so on.

[20] An example of useful metadata is the information provided in e-mail messages which contains simple metadata in attribute values in the mail header. Programs can use this data to provide threaded discussion archives, searches for names or topics and so on. s

[21] Computer programs cannot, of course, on their own carry out any "intelligent" analysis on a give text. For example tell the difference between the personal pronoun *I* and the numeral *I* unless rules have been set. This information can be added internally in the actual document or a set of externally defined rules can be applied to a text.

[22] I use this expression inspired by Willard McCarty (2002) who is discussing the kind of experimental and testing practice humanities computing is.

[23] This is further elaborated in e.g. Alexander& McKenzie, 1998; Blanton et al., 1998, Langerth Zetterman, 2001, Strömdahl & Langerth Zetterman (in press).

[24] See e.g. Iivonen & White, 2001 who studied differences in how users from different cultural groups searched for information.

[25] Cf. Laurillard, 1993; 2000.

help researchers and teachers to use and reuse research and teaching material – the actual content – without proposing a special teaching method or instructional technique.[26]

Therefore it is a need for more activities concerning how teachers and students might create, manage, share and reuse content. To explore uses of digital archives and electronic texts is relevant to education because it reveals opportunities to an alternative to handbooks and predefined teaching methods for teachers in higher education. A question here is: is it possible to study the use and organization of content in teaching without having a special method in mind? Yes, I believe so, because advocating a flexible use of digital content in higher education is not the same as assigning *one* special teaching method. On the contrary, it rather is to have a standpoint that teaching in higher education could some times benefit from a flexible access to and use of digital content, whatever the purpose of uses might be.[27]

These Issues can also be posed as; how might digital archives be designed allowing teachers and students flexible modes of access and retrieve and use of the content therein?

If the content is separated from its presentational forms one can easily imagine how the same information could be used in many different occasions and for several different purposes in research and education. Similarly, if the content is digitally produced and stored with no predestined use or predefined teaching method the same content could be used and reused in different teaching and learning situations. Consequently this also means that the same original content might be personalized, accessed and manipulated by teachers and students at different educational levels or different disciplines and subjects.

In domains such as distance higher education where students are dispersed, sometimes far away from university libraries, their access to relevant content is most crucial. It is also important to note that learning does not end with the completion of formal education. On the contrary, policy in many developed countries focuses on the need for lifelong learning and recurrent education. Much of the learning occurring throughout life is supported by other means than reading textbooks or formal training. For the individual the affiliation to different kinds of communities is crucial in order to obtain learning opportunities. Today and even more in the future such affiliation often mean access not only to the members of the community but also to the community's shared resources on the Internet. Therefore, it is a problem that the so called eLearning markets often provide proprietary solutions for content management and communication.[28]

Another aspect of proprietary solutions is when educational materials are produced solely for educational uses[29] and when tools become specialized for schooling purposes. Then, educational practices distance the tools form other kind of uses simply by labelling them educational tools. Instead, the development of tools and methods aimed for research and education might be done the contexts in which they are being used. Jerome McGann, a

[26] I am convinced that there are some studies investigating the use of content without advocating a special instructional method and I will hopefully find more of them during my studies. One example is Julia Flanders (2002) study of uses of the digital archive "Women Writers Online". See also Susan Hockey 1999 & 2000 for lucid and useful introductions to technical issues of document management in scholarly practices.

[27] Cf. Broady (2001). Digitala arkiv och portföljer", pp. 11-16 i *IT i skolan - mirakelmedicin eller sockerpiller?* Rapport 45/2001. Stockholm: IT-kommissionen.

[28] The rather stereotyped kind of flexibility of the "Elearning market" provides has been criticized by e.g. Langerth-Zetterman & Lindblad, 2001. This on-sided flexibility provides flexibility in time, place and pace but not for e.g. content modularity or interoperability.

[29] Illich use the term "monopolized" when talking about educational materials, in his article "A Special Supplement: Education Without School: How It Can Be Done." The New York Review of Books. January 7, 1971.

humanist scholar doing research in literary and textual theory, brings this issue right to the point:

> We read, we write, we think in textual condition. Because this is true, the new information and media technologies go to the core of our work. As humane scholars we should not leave the development of these tools, which includes their introduction into our institutions, to administrators, systems analysts, and electronic engineers.[30]

An important design principle for digital content would be that the structure of the digital module in archives should be separated from the structure of the actual course given. The digital content modules should be equipped with metadata and when possible apply to relevant international standards.[31] By separating what is to be learnt from how it is taught, one and the same content or archive might be used in different courses by means of different filtering and presentation, and by different target groups, from freshmen to specialists. For a specific course the teacher might propose the students certain paths through the archive and a certain subset of content modules to be added to their digital portfolios.

Digital portfolios[32] might serve as a personalized content collection acquired by, for example, students during the course of their years at the university. A portfolio might include the student's own or peer students annotations, papers and project presentations, courseware and reference literature, material created by their teachers, test and examination results, copies of or links to various resources. It is also crucial that student's are offered course content, teachers' commentaries and guidelines, test results and other relevant material in portable modularized formats, suitable to be incorporated into their portfolios or personal digital archives and to be reused for various purposes that may not be foreseen by the teacher.[33] Thus, the personal archive or the portfolio might serve several purposes: depository of material for personal use or to be shared with other students or with teachers, documentation of the progress of the studies, reference points in the career planning.[34] When the student leaves the university it might be useful in future professional activities. When applying for a job it might contain items to be presented to an employer.

Many issues concerning the use of portfolios in education has been dealt with at numerous occasions elsewhere.[35] Studies about portfolios are often focused on how portfolios can be used for e.g. student collections, student reflection (between research based knowledge and their own learning, reflect on previous experiences), and gathering evidence of professional development.[36] There are also examples of studies for cross-disciplinary and collaborative uses of portfolios.[37]

Although the use of digital portfolios is an important aspect I will not elaborate *the concept of portfolios* in higher education per se in my PhD project since it is not the purpose of my study. At times there might be points in common in my forthcoming studies, especially when uses of personal digital archives or shared use of some information components. But otherwise I will not study any aspect of for example, how portfolios might help students in

[30] McGann, 1998:609.
[31] See e.g. Paulsson, 2003 and Wiley, 2000.
[32] See e.g. Sjunnesson, 2001 for an overview of digital portfolios in teacher education.
[33] A study of using digital portfolios in courses in teacher education with no predefined method in mind will elaborated in Gustafsson & Langerth Zetterman (working paper in process).
[34] This is one of the purposes with the Standardized Content Management System, SCAM: which is an archive system for storing and distributing learning components. See e.g. Paulsson, 2002.
[35] E.g. by McNair & Galanouli, 2002, Meyer & Tusin, 1999 and Sjunnesson, 2001.
[36] Cf. Sjunnesson, 2001 and Meyer & Tusin, 1999.
[37] Fils, Taber, Takle & Sorensen, 2000 and Tolsby (in press).

their learning or how they might be reflecting on the content by using a portfolio. The shift from proprietary teaching materials and tools to a Web based and multipurpose usage of different kinds of resources, such as a digital archive, is, however, to the purpose. For example, the notion of "resource-based teaching" might be useful to elaborate since it seems to signify a convergence of pedagogy and research, by offering solutions where resources might be separated from their manifold and involves highly changeable uses. [38]

On the relation between research and educational practices – scholarly uses of digital archives

Above I have argued that there is a need for archives of content modules on the web - easy to share, to navigate, and to combine and reuse in new contexts. If such archives are to be beneficial in various educational settings, and if they are to stimulate freedom of choice in higher education, it is crucial that the content is adapted to emerging international agreements and standards on mark-up languages and metadata. Else, the content will be locked into proprietary platforms and applications. My aim is therefore grounded in a standpoint on the benefits of multiple uses of digital content and flexible content design, of e.g. original sources and research material in higher education, as an important alternative to teaching materials which are predefined and predestined for specific courses or situations. Then, the question here is; how might teachers and students benefit from emerging mark-up and metadata recommendations and standards?

My standpoint incorporates an notion of student learning and working which is more alike research.[39] That is: to provide opportunities for students to find "[...] a closer approximation of the thing itself"[40]. While confronting a wholly or partially unknown "body of knowledge" [41] is a challenge in reality for many teachers and researchers, the situation is opposite for many students in higher education, a challenge which is effectively avoided by an extensive emphasis on syllabi. I want to emphasize that I am not advocating a "teacher-less education" or something alike. On the contrary, the teacher's role should be strengthened by giving opportunities to teach in a context reminding of research work. The teacher might provide judicious points of departure, suggest possible strategies, and guide students in their exploration of content and materials.

In a situation when educational systems and resources proliferate, the need for standardization becomes apparent. Like in any standard driven initiatives, standardization applied to educational settings and research is focused towards enabling, reusing and interoperability among heterogeneous computer platforms and software systems. To achieve this, consensus is needed on a number of issues e.g. architectures, services, protocols, data models, classification schemes and interoperability between different standards - whether they are de facto or established by the ISO standardisation organisation. This is an active, continuously evolving process that will last for years to come and a complex process, which occurs at several levels and is supported by many different related and standalone initiatives.

Among the most important general standardization efforts when it comes to mark-up languages and metadata, which has to serve as the prerequisite for research and educational practices described in this paper, are SGML (Standard Generalized Markup Language, ISO

[38] See e.g. Chambers, 2000, Spaeth & Cameron, 2000.
[39] Cf. Broady, 1995; 1997; 2001, Bourdieu 1996a&b, Illich, 1970 & Flanders, 2002. Although inspiration has been gained through readings of disparate texts and authors I here focuses discussions on the *usage of content or teaching material* although the issue is somewhat differently argued and discussed
[40] Flanders, 2002:50.
[41] Ibid.

8879:1986)[42], XML (Extensible Markup Language)[43], RDF (Resource Description Framework), Dublin Core[44] and Topic Maps[45]

Within the domain of the so called instructional technologies there is an important international project aiming at creating de facto standards for educational systems and resources is project IMS (Instructional Management Systems)[46], started in 1997 and since then engaging a large number of major educational institutions as well as software companies. Its aim is to develop and promote open source specifications for many domains of distributed online education, including specifications for organization and delivery of educational content. Another example is the ARIADNE project within the European Unions 4th Framework R&D Program. This project focused on development of tools and methods for producing, managing and reusing pedagogical elements. Results from ARIADNE are e.g. recommendations for educational metadata that follows a certain structure, were elements of pedagogy and semantics are taken into account. The project is thus an attempt to surpass general and technical information about the resources.[47]The educational metadata schema LOM (Learning Object Metadata)[48] aims at helping teachers to specify and find multimedia materials to suit their pedagogical needs or preferences, by facilitating content sharing and reuse among educational professionals. The "learning object" concept is grounded in the object oriented branch of computer science and the engineering community. [49] The idea with learning objects is to produce, (re)use and combine small chunks or components of *instructional* materials.[50] Although there are several positions taken by different communities within the instructional technology area, most of them are commonalities: small, digital, reusable, educational resources.[51]

Potentially these kinds of international standardization efforts will be useful for teachers and students. Provided that courseware and other educational resources are made available in modular shape and adequately marked-up, one and the same content might be combined and reused for different audiences with different needs in different educational settings.[52] Teachers and students might be encouraged to find personal ways of learning and less inclined to use pre-constructed syllabi and selections of course content, as mentioned earlier in this paper.

---

[42] SGML is an ISO-standard since 1986. The seminal work is Charles F. Goldfarb, *The SGML Handbook*, Clarendon Press, Oxford 1990. For more information on SGML see e.g. http://www.oasis-open.org/cover/. The Swedish SGML user organisation is found at http://www.sgml.a.se/

[43] XML is developed by W3C (the World Wide Web Consortium). Version 1 was adopted in February 1998. See http://www.w3.org/XML/

[44] The Dublin Core Metadata Initiative (DCMI) is an organization working towards the adoption of interoperable metadata standards as a supplement for existing methods for searching and indexing Web-based metadata. See http://dublincore.org/

[45] Topic Maps (ISO/IEC 1999:13250) is a new standard for the description of knowledge structures and the relations of these structures to information resources.

[46] The Instructional Management Systems, IMS, consist of academic, non-profit, corporate, and government organizations. See www.imsproject.org/ for further information.

[47] See: http://ariadne.unil.ch for details.

[48] LOM is specified by a working group of the IEEE Learning Technologies Standardization Committee. See http//:ltsc.ieee.org/wg12/.

[49] The notion of learning objects in instructional design is promoted by Learning Technology Standards Committee (LTSC), see, http://ltsc.ieee.org

[50] See Wiley, 2000 for a definition and overview on learning objects. See also Paulsson (2003) and Sjunnesson (2003).

[51] Other approaches (apart from the already mentioned learning objects) are: educational software components, sharable content objects, knowledge objects or educational objects (Wiley, 2000).

[52] Paulsson (2003) provide a useful framework for a exploring and using component based teaching environment. s

Furthermore, the move away from proprietary solutions and towards open standards will offer better opportunities for teachers and students to benefit from existing digital archives and other resources, which are today mainly used by research communities and other specialists.

However, today the benefit of appropriate mark-up is not evident to most users in the educational sector. On the contrary, teachers and students are prone to accept courseware based on ad-hoc categorization, redundant information, proprietary applications and formats such as PDF, non-modularized solutions, and rudimentary mark-up (typically raw HTML). This attitude is quit understandable, for several reasons. The courseware publishing houses are reluctant to deliver content in portable formats since they - as well as the software producers - prefer to keep the costumers tied to the vendors' proprietary solutions.

A second reason is that teachers are not used to separate what is taught from how it is learnt, that is to separate content from its presentation forms. Therefore, they do not spontaneously applause the opportunities to utilize different filtering and presentation tools in order to use the same digital archive for different courses and/or for different student groups. They do not consider the possibility to accumulate content into archives through which the students might be suggested paths or from which subsets of content modules might be incorporated into the student's portfolios - or even produced by the students themselves. Instead many teachers tend to regard content as hardwired into the syllabus of certain courses.[53]

A third reason why teachers are lagging behind many research disciplines and industrial sectors is the shortage of appropriate and easy-to-use tools allowing teachers and students to access portable resources and to create their own content archives or portfolios.[54] The creation of structured marked-up and metadata-enriched content still requires skills that are fairly rare within the educational sector. Even if XML editors are available the mark-up procedures still are too non-transparent and tedious for most users. Even if the current versions of ordinary web browsers have some support for XML, most teachers do not have any clue about how to make use of those facilities.

In order to find out how teachers and students might benefit from emerging mark-up and metadata recommendations and standards there is a need for both better tools and more developed methods. The proposed PhD project might be a modest contribution to this end. Both by investigating how existing digital archives are used and both by the application of a few tools in the digital archive which is developed and maintained within the research group where I participate; Digital Literature at Uppsala University.[55]

Content design in research communities

In order to find inspiration for more general educational applications, I will also explore how digital archives are in fact used in some research communities.

The emerging development of multi-disciplinary fields and heterogeneous audiences at universities makes communities even more complicated both in terms of the disparate nature of shared digital resources and because of the more heterogeneous background of the participants - as in bioinformatics where PhD students might have their major in computing science, mathematics, biology, chemistry or medicine.

[53] See e.g. Broady, 2001 and Flanders, 2002.
[54] See Sjunnesson, 2003 for an overview of metadata for learning objects on the Web.
[55] See http://www.skeptron.ilu.uu.se/broady/dl/index.htm for further information.

I have argued elsewhere in this paper that by using modularized and structured digital archives based on international standards teachers and students could get distributed access and opportunities to build personal paths for their teaching and learning. Of course the uses of technologies are having a *different* kind of impact in *different* communities.[56] To explore and study the different sorts of impact in different settings might contribute to a better understanding in how to design content with a flexible purpose in mind. I will argue that a total absence interpretation or thinking of future purposes in content design might not be possible although there are many possibilities to design content allowing for a much greater flexibility than textbooks or on-line teaching materials usually does. It might also be of interest what kind of instances and the amount of necessity scholars might have to take into account according to the immanent structures and constraints within scientific disciplines when practicing content design, e.g. preparing and encoding electronic texts. .

Teachers and students will also have more choices and a better overview over the resources and the context into which these resources are inserted. To some extent these opportunities are realized in certain university disciplines and environments. Therefore I propose further exploration of some communities where modularized content archives are heavily used. Two such communities, which I intend to approach by means of case studies, are bioinformatics and the TEI community in the humanities. By this research design - that is "test beds" in the most advanced natural sciences on the one hand and in the humanities on the other – I hope to reach some transparency in our understanding on how digital archives are managed and utilized today within research communities, and possibly in the future within some educational environments.

Such communities are groups of individuals who work, learn and socialize together sharing insights and develop shared knowledge as a consequence of participation. Communities thus evolve, develop and merge around shared interests and expertise. Recent research has highlighted the importance of tacit group knowledge within communities, i.e. knowledge not held by individual members but reflected in the artefacts created and shared by the community.[57]

In a university setting, the members of such communities are researchers, teachers, students, and administrators. Their shared digital resources are any object or location that has a unique identifier and can be digitally stored, accessed and distributed via a global network or a local area network. A resource might thus be web pages, web sites, course modules, lecture notes, reports, papers, databases or students digital portfolios.[58]

Another aspect is that information technology also allows new and innovative uses of elderly and timely communication patterns of e.g. the output of research. Scientific communities develop and manage their own information nodes on the Web either to speed communication in time and across space or to make fragile, unavailable and/or large textual materials accessible for research and manipulation. As for example in some scientific community, this has led to such innovations as the establishment of preprint archives in high-energy physics.[59] On the other hand a lot of the progress in the emerging mark-up and metadata recommendations and standards are taking place in interdisciplinary research communities. In a field, such as genomics, they are building massive databases which require information management and computer science specialists, forming an academic domain of its own - bioinformatics.

[56] Cf. Knorr Cetina, 1999.
[57] See e.g. Cook and Brown, 1999 and Schön, 1988, who in this article describes how designers share models that serve as holding environments for ideas that need not or cannot be articulated.
[58] See Sjunnesson, 2003 for a discussion about resources as learning objects.
[59] See http://arXiv.org

Another example is humanities in computing, working at the intersection of computing, arts and humanities. In the humanities computing community, there is both a pragmatic focus on application of computing in scholarship and teaching, but also a theoretical focus on sociological and epistemological issues. One position posed by Willard McCarty is that humanities computing are an academic field of its own concerned with theoretical and epistemological issues like "how we know what we know". On the other hand there is Lou Burnard, who has a different standpoint more related to the inherent logic and practices in different academic practices. He is posing the question: 'What use is this technology to my academic concerns? A similar view is expressed by Susan Hockey.[60]

One prominent example in the humanities domain is the previously mentioned TEI community. TEI was originally developed as an application of SGML - an interchange language for textual data. This interdisciplinary effort has been very successful and is now widely recognised as the standard format for scholarly text encoding and textual interchange.[61] The TEI community, have had an impact on the emerging mark-up initiatives far beyond its own scholarly domain by the contribution to the development of XML, the dominant standard for information interchange format for web services.[62]

Besides, text encoding allows texts to be treated as research tools in themselves. That is, digital texts lend themselves to much more than access, retrieval and reading. Encoded texts can help scholars do other kinds of research work. Full-text resources offer, at least three, clear benefits; (i) a simple provision and access to of otherwise scarce texts, (ii) advanced search opportunities either to identify particular motifs or words or to establish their absence in certain texts, and (iii) the ability to collate different editions of the same work for variants or to identify editorial changes.[63]

In a similar interdisciplinary manner, yet with a totally different scope and aim, other communities are working on the development of external meta-data recommendations and standards for the interchange of e.g. educational information between different platforms and contexts. In some of these communities, such as the IMS project, they tend at a first glance; to have a somewhat unproblematic view of what can be considered as elements of (pedagogical) data. Another example is the metadata recommendation Dublin Core Metadata Initiative, mentioned above, is an initiative and a standard for cataloguing internet resources and where there has been a tremendous will to work towards consensus on a simpler solution of interoperability between different domains. Thus, by using a set of minimal requirements of how to describe resources the Dublin Core are striving for making information applicable, understood and useful in many different contexts.

Communities of scientists and researchers concerned and interested in issues on electronic text, metadata and interoperability are thus created around the globe with access to essentially the same information and with fewer of the sociological or physical barriers that previously existed. But, the limited use that many scholars (others than the humanities computing community) have made thus far of encoded texts is not due to insularity but rather unawareness of how to use the texts (digital archives) and unawareness of the many opportunities for research offered through encoding.[64]

What I have tried to pinpoint above is that the advent and the usage of digital archives, web based repositories and databases have precipitated the blurring of not only geographical

[60] Cf. Burnard (2001), McCarty (1998) and Hockey (2000).
[61] See e.g. DeRose, 1999, Hockey, 2000, Renear, 1999.
[62] Important TEI members as Michael Speerberg-McQeen, has been a co-editor of the W3C XML specification. See DeRose, 1999 for a summary of TEI's many contributions to XML.
[63] These benefits are discussed by e.g. Burnard (2002), Hockey (2000) and Sutherland (1997).
[64] Brockman, Neumann, Palmer & Tidline, 2001.

boundaries in research communities but also creating new kinds of communities and fields by working and exploring the same, or at least similar, kind of interests. Yet, few social scientists have investigated the ways in which these information databases and digital archives are being used within and across research communities. I am curious in many ways: How do the researchers work in these communities? What kind of principles and interests structures their orientations? How do they handle the complexity of their information resources, be it textual or other kinds of data such as sequences, annotations, analysis results, database links, graphical images, etc.?

A more pragmatic aim is to test tools that can help community members to share and retrieve communal artefacts by being "tuned" to the community they are intended to serve. One motive for adopting this approach is that communities have particular characteristics, which e.g. are revealed in their way of doing things and the domain-dependant use of concepts and topics. Here I expect to be able to profit from ongoing standardizations efforts in the domain "topic maps".[65] This standard is aiming to express concepts and topics in such a way that they can be presented and shared on the web. Topics and topic associations build a structure semantic network above the resources of information allowing description and retrieval of data similar of using indexes and different hierarchies to find specific topics. The topic maps enthusiasts believe that we are approaching a significant transformation of the web, from presentation of information to representation of knowledge.

Some methodological and theoretical considerations

Before proceeding to the section of proposed methods I will present some initial notes on previous research and notes on methodology that might be relevant to my project. I am still in a process of trying to understanding several possibilities or perspectives that might be useful for this particular project and I am therefore a bit uncertain of how to tackle the research problem. I would like to emphasize though, that through the process in my attempts of constructing the research object I am constantly trying to break with the common-sense problem definitions and the initial practical sense.[66] In my understanding, Bourdieu suggests that we can figure out the sense of reality and that everything has a cause or a reason but it might not be what we thought it would be from the beginning when we used the everyday glasses.[67]

I am not sure I am, at this stage anyway, able to reach beyond these limitations - but the questions, methods and analysis outlined in this proposal is an attempt to understand and overcome this kind of limited initial practical knowledge. I also want to emphasize that the experimental practices I want to explore are fairly new to me so the practical sense I bring into this context mainly originates from other kinds of practices.

How do researchers work with their tools and methods?

Previous research about what researchers actually do in their laboratories, the production of knowledge and scientific work has been undertaken by quite a few; such as researchers within the Sociology of Science, a perspective including several and contrasting views. That scientific and other beliefs are largely determined by social causes (e.g. Bloor and Barnes), or scientific facts are constituted by social interactions (e.g. Knorr Cetina, Latour & Pickering) and that the empirical content of scientific statements is perpetually open to

[65] See e.g. http://www.ontopia.net for further information on Topic Maps.
[66] Cf. Bourdieu, Chamboredon & Passeron, 1991.
[67] Cf. Bourdieu, 1996a:40-45.

re-negotiation (e.g. Collins) and other sociologists and philosophers delivering critique on one or all of the other perspectives (e.g. Bourdieu & Hacking). In being a novice in the studying the sociology of science I will here concentrate on just a few glimpses in this vast and complex area of research. I have to do more readings in this area in order to comprehend and understand the different perspectives within and what they know about how scientists work and how knowledge is made.

Karin Knorr Cetina's *Epistemic Cultures* published in 1999 is an empirical study involving more than ten years of observations at two different laboratories, one in High Energy Physics (HEP)[68] and one in Molecular Biology[69]. Her thesis is that the study of scientific fields exhibit distinct "epistemic cultures" – a thesis that is strengthened when she shows that *different* scientific fields exhibit *different* epistemic cultures.

I intend to try what Knorr Cetina suggests in her book: that other should use the patterns she discovers in her research on high energy physics and molecular biology as "[..]templates against which to explore distinctive features of other expert domains." [70] My intention is to use her findings in studies of other epistemic cultures, i.e. in this case equivalent to what I choose to call experimental practices.

Knorr Cetina put forward that in her studies, especially in high energy physics, the collaborative process of "unfolding" characterises decisions and reveals procedures and features illustrating, what she calls, liminal phenomena –knowledge about the limitations of knowledge. Because there is no body of comparative results, physicists use this liminal approach, arguing that they must constantly test, calibrate and question the kinds of results they are getting: this kind of activity she is calling the "care of the self."[71]

Among the most salient features found through observing the HEP experiments are:

- the size and complexity of the machinery (the laboratory is a complex experimental setting)
- the size of the collaboration (involves thousands of people from time to time)
- the long duration of experiments (experiments typically take several years)
- the unstable structure of collaboration (due to size and long duration of the experiments and participants come and go)
- the physical separation of participants[72]

Microbiologists, on the other hand, have their workshops and labs to turn to for results and comparisons; consequently they are more focused on *results* than the actual instruments. They also more freely turn to run "blind" variations within an experiment to see what the results will be. Microbiology decisions are more hierarchical and this structure differs from the collaborative efforts of HEP experiments where all the complex machinery and measurements necessitates human coordination. One level in molecular biology consists of

[68] Knorr Cetina's first and most extensive case is high energy physics(HEP); especially experiments done between 1987 and 1996 at the European Centre for Nuclear Research (CERN) in a huge laboratory (27kilometers around) of Large Electron Positron Collider, located on the border of France and Switzerland. This collider is now replaced by Large Hadron Collider (LHC) and very large detector called ATLAS (44 m wide, 22m high). The experiments in HEP involves more than hundreds of scientists – from time to time more than thousand physicists. Knorr Cetina 1999:15-25.
[69] The study of Molecular Biology included the Max Planck Institute group working on cellular biology in Göttingen and the group of scientists observed varied from eight to thirty during the years. The observations started in 1984 and were at 1999 still ongoing although involving other groups. Ibid:18-19.
[70] Knorr Cetina, 1999:252
[71] ibid. pp.55-61.
[72] ibid. pp. 159-191.

individual researchers each of them working on their own project and the other level consists of the whole laboratory usually managed by a single director. Knorr Cetina argues that the individual nature of the level has important theoretical implications.

> This is perhaps molecular biology's first most important difference from experimental high energy physics: in the molecular biology laboratory, the person remains the epistemic subject. […]The laboratory, experimentation, procedures, and objects obtain their identity through individuals. The individual scientist is their intermediary—their organizing principle in the flesh, to whom all things revert.[73]

What about the research communities I want to study then? Firstly, bioinformatics is concerned with the gathering, analysis, and exploitation of data and is one example of a research community at the interface of biology, medicine, mathematics, and computer science. Biological data is not only generated in overwhelming amounts today, it is also of a widely disparate nature. The development of methods to integrate and exploit the various data sources requires competence in computer and information science. The scientific knowledge, however, is biological in nature and a basic understanding and appreciation of biology and biomedicine is crucial for any integration methods to become successful.[74]

For much of the time since the inception of humanities computing in the 1940s[75] until the late 1980s humanities computing was largely the province of individual scholars doing "things" on their own[76] for research but also for teaching. Scholars created electronic text(s) and then subjected those texts to different kinds of analyses, with a concordance program or perhaps some other custom programs as well. The electronic texts, created as the by-products of these individual research projects, usually reflected the theoretical viewpoints of the scholars carrying out those projects. Some of these resources have found their way into archives or digital libraries. Besides, the issue of reusability did not gain any significance until the late 1980s. It then became generally accepted among many researchers that a common encoding format would make it much easier for researchers to exchange and reuse the electronic texts they where producing.[77] After the proposal in 1987 the methodological commons[78] is the centre of this multidisciplinary and heavily technique depending practice. The humanities computing community consists of members from all of the humanities but also to some extent from closely related social sciences.[79]

What then is a laboratory and how do they do experiments in the humanities computing and bioinformatics? I observe a striking, non-trivial resemblance between humanities computing, bioinformatics with some of the salient features of HEP (see above) they are: data-centred, heavily equipment and machinery -orientated activities that centrally involve some kind of modelling and they tend to be collaborative endeavours. Then, the lab might be the setting in which the researchers use this machinery, since the computers with the accompanying applications, tools and methods are the environment in which the experimenting take place. As Knorr Cetina points out "it [the computer] provides its own

[73] Knorr Cetina, 1999:217.
[74] Andersson, Langerth Zetterman & Strömdahl, 2001.
[75] Susan Hockey, Willard McCarty and other denotes the beginning of humanities computing to the 1940s when a Jesuit scholar Italy, Father Roberto Busa, known as the founder of Literary and Linguistic Computing, created an exhaustive concordance of the writings of St Thomas Aquinas, the *Index Thomisticus.* The first volume of the *Index* appeared 1974 and was published on CD-ROM 1994.
[76] The notion of the "lone scholar" is currently discussed in the Humanist listserv.
[77] The TEI project was formed to create an encoding format, Hockey, 2000.
[78] See McCarty, 2002.
[79] A useful map picturing what kind of disciplines and activities involved in humanities computing, proposed by McCarty, can be found at: http://www.kcl.ac.uk/humanities/cch/wlm/essays/encyc/figure1.html

test-bench environment" when talking about doing experiment by using computer simulations.[80]

I am also curious if one might find similarities between the molecular biology and bioinformatics. What I need to do then, is to ask where a kinship lies and what kind of assumptions one might be able to make from such kinships? On analyse instrument which might be useful here is Joan Fujimura's uses of concepts like "boundary objects" and "standardised packages".[81] These concepts are means of describing and analysing how collective action is managed across different social worlds to achieve enough agreement, and not necessarily consensus, in order manage production and to get the work done but also to find a temporary stabilisation for handling "facts". The concepts might be used for analysing and understanding how researchers manage to collaborate, to translate different views and how they work towards similar goals, not needing to have the same kind of understanding or knowledge of the objects under study. Fujimura describes how the two objects have developed through different set of studies which have in common the heterogeneity among the social worlds involved. The concept of "standardised packages" is about how a community (she use the definition of cancer as an example) have a common ground of scientific theories and a standardised set of technologies for collaborative work and stabilisation of "facts" – it is used to define the conceptual and technical universe in which the research can take place. This concept, might be seen as a "grey box", serving as interfaces between several contexts, combining several boundary objects (in her example gene, cancer etc.) and is less abstract than a boundary object. A boundary object then, is an object "[…] having different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation.[82] A boundary object in my studies might be "electronic texts" or "digital archives" and a standardised package might be "humanities computing" or "higher education". I think these instruments interesting and possible useful for my purposes but I need to look further into the underlying assumptions.

On the other hand there is Hacking who discuss the diversity of scientific methods - there is not one scientific method because the methods are as many as there is topics and sciences, and that questions of method arise in that context.[83] What Hacking and also Knorr Cetina are suggesting is that experimental practices have its own epistemological agenda which need not to be introduced by a priori theories.

McCarty also argues, in his research agenda for the humanities computing community, that a useful tool for analysing the practices of humanities computing might be Hacking's model of experimental sciences where the investigator make hypothetical entities real by learning how to manipulate them.[84] This is similar to the notion of liminal knowledge I mentioned above. And it might be useful when I am analysing data related to exploring and comparing the salient features inherent and embedded in the practices I want to explore.

Finally I return to Knorr Cetina who claims that by using one epistemic culture one may look at one science through another and that the comparative optics might bring out not the essential features but the differences which otherwise would be hidden.[85] These differentiated features (i.e. the revealed differences) might even be the only tractable elements available to us. Would it be possible to use these differentiated features for revealing relational and distributed structures of some experimental practices?

[80] Knorr Cetina 1999:34.
[81] Joan Fujimura 1992:168-211.
[82] Fujimura 1992:173 (citing Star & Griesemer, 1989).
[83] Hacking, 1999:198.
[84] Cf. Hacking, 1992:44-50.
[85] Knorr Cetina, 1999:4

Studying scholarly work in the humanities

In another and much more modest study, done a few years ago, the investigators are drawing conclusions about the work and the research process through extensive observation of selected humanities scholars.[86] The essential findings in this study were that humanities scholars have adapted well to rapid technical change and that they are to a reasonable extent, able to harness information technologies to somewhat traditional research functions. Such functions include, for example, keeping abreast of a broad secondary literature surrounding their fields of inquiry. It also includes the process of locating, acquiring access to, and using primary resources that are relevant to a particular area of investigation. The scholars themselves are also developing personal libraries which they feel are enriching and reinforcing their scholarship. The investigators also claim that these patterns of research practice offer valid guidelines for digital content archives such as research libraries because these patterns disclose the actual processes and the essential role of texts within the humanities. This position is grounded elsewhere,[87] e.g. by editorial theorists advocating the text as a centre for (almost) everything humanist scholars do.

The investigators also emphasize, without mentioning any particular study, that previous work on the research practices of humanities scholars has differentiated research work from activities of keeping the subject current and teaching preparation. However, in this study it was clear that the two latter activities are important complements to research and that the three types or work are, according to the scholars themselves, inextricable and complementary. While they in the study, examined the day-to-day practices of the studied scholars, their activities were documented as well as the resources involved – i.e. to trace and collect the overall scope of scholarly work as such. Besides, and this I believe is important, while in this study the investigators thought they focused on the actual research process, the results they presents might not be exclusive to research work, on the contrary, it involves the teaching and other activities closely related to the research practice.[88]

The study outlined above also contains references to other studies and surveys which I have not had the opportunity to read yet. These studies, I hope, might be useful to me in my understanding of my own case studies and as complement to other studies of research practices mainly focusing on the sciences rather than on researchers in the humanities or social sciences.

Methods- three steps in the research

The methods and the procedures I intend to use and described below suggest how I might tackle my aim and the questions posed in this project. I am aware of that the proposed three steps altogether might be too much for me to accomplish in this Ph D project, therefore I am hoping for suggestions and commentaries in what way and why the different steps (or any of the steps) might be useful in my forthcoming studies.

One way of improve the exploration of technology uses in educational settings is to focus on the use of tools providing opportunities related to teaching and learning that could not be provided otherwise.[89] Another (although quite obvious) conclusion is that we need to know more about what good practice is, if we are to offer the best possible for a particular

[86] Brockman, W.S., Neumann, L., Plamer, C. & Tidline, T. (2001) "Scholarly work in the humanities and the evolving information environment". This study used a variety of methods (such as interviews, document analysis and workspace observation) to study the 33 participants from different areas in the humanities, mainly from language departments, at University of Illinois and the University of Chicago.
[87] Foremost by Jerome McGann - a proponent of editorial theory.
[88] ibid. p. 6-28.
[89] Cf. Alexander& McKenizie, 1998.

group of learners engaged in learning a particular content.[90] In the proposed project I will pay attention to some of these factors, by a systematic exploration of some research communities with a considerable knowledge and experience of both good, and surely bad, practice in the design and uses of digital content archives.

As stated above, my aim is to explore and study how we can make information applicable for different purposes and uses in research and educational practices. Methods to be used are studies of documents, interviews, participant observation, and user studies on the application of certain tools and methods. The research communities I have chosen fore the case studies are the humanities computing community and the bioinformatics community.

What constitutes the activities of humanities computing and the bioinformatics and how might it be studied? What are the characteristics of the objects of study, the knowledge production and the relationship of scholars and the content they produce? What kind of attributes or characteristics could be appropriate to use when exploring the communities?

The exploration should be a new construction of already existing material experienced and organised in a way, not similar to a spontaneous experience or common ways of classifying. The basic assumption then will be to re-organise, re-interpret and re-classify the reality to figure out whether and why these practices might be what they are.[91]

The variety of methods might be necessary in order to understand how tools and methods can be used to make content retrieval, access, use and reuse transparent for research as well as education. On the following pages I am elaborating and explaining the proposed methodologies I intend to use while exploring issues posed in my questions.

First step: an overview of how content design and metadata schemes are used for research and for educational purposes

This initial step will result in a systematic overview of some selected pertinent and ongoing content design and metadata schemes[92] for research and education.

The aim here is literally to draw a map of initiatives which have implemented standardization and metadata guidelines for pedagogical and scholarly uses and are providing opportunities to prepare and use full text digital archives in the humanities and the social sciences. The attributes I would need to picture this territory are not evident in any way and that's why I intend to explore this territory without an extensive a-priori classification apparatus.

This overview will be useful for my project in understanding both the specific but also the wider context of digital archives when carrying out the case studies of some research communities (the second step in the research, see p. 24). During my search for relevant background material and projects I have found that metadata initiatives are on an opposing relative scale to each other in a number of ways, many of them equally important. This is not surprising since different domains of course put forward their point of view. Many initiatives found on the web are mostly concerned with the adding of external metadata, often implemented in XML (e.g. the Dublin Core or the Resource Description Framework, RDF) while others are concerned with embedded internal metadata, mostly implemented in SGML or XML. Some of these initiatives, such as topic maps, are established ISO standards and others are guidelines (i.e. the TEI DTD) serving as de facto standards.

[90] Alexander & Blight, 1996:1.
[91] Cf. Bourdieu, 1996a:40-45
[92] From here I will use the term, initiative or project in analogy with examples of individual and organisational practices where they have implemented content design and metadata schemes in digital archives.

I have encountered many difficulties designing how to do this overview and to start with: where should I draw the boundaries of which initiatives to include in my overview? I will mainly collect the data relevant to this overview through the web so the first criteria will be that the digital archive is available on the web. Other issues I have to consider are the levels of manipulation of content; spanning all the way from just delivering texts on the web with no particular metadata added, through adding descriptive external metadata or internal metadata aimed for interpretation, manipulation and analysis.

There are also initiatives which is more orientated towards the digital library domain and thus mostly concerned with preservation issues or digital publishing and scholarly journals – issues that are not in any way unimportant for my aim but not in the scope for this overview.[93] On the contrary, all these different kinds of overviews and surveys have been important in terms of getting a grip with the state of the art in digital archives for the humanities and social sciences.

This first step will be accomplished by two interrelated studies resulting in three articles addressing the explicit strategies of the selected project and/or initiatives responsible for the digital archive(s). The first article will be a detailed elaboration of the methods with some initial results from the overview and the second article will be the actual map over the digital archives included in the survey. The third article will be a retrospective study of the TEI community.

A typographic exploration - revealing relations and structures

Overviews and analysis of semantic web activities and learning technology standards have been and are being done elsewhere[94] and many projects are still in the developing phase and have not yet actually launched a full text digital archive.[95] I am choosing to focus my survey on initiatives which are in fact using the TEI DTD guidelines or other comparable text mark-up languages (DTD:s) enabling semantic interoperability. According to several proponents the Text Encoding Initiative, founded in 1987, is the single most important multidisciplinary collaborative project in humanities computing up to this date.[96] This is just one good reason to explore the results and the strategies of the collaboration and the work done in this environment.

In order to find the relevant projects to study I will use existing directories and gateways such as the TEI Consortium's web pages currently including 107 projects using the TEI guidelines worldwide[97] and the Humbul Humanities Hub[98] which is a gateway to online resources in the humanities and the social sciences. These gateways are updated on a regular basis and will thus ensure a potential actuality of initiatives to select for analyse.

The already existing comprehensive overviews and surveys describing projects and initiatives using different metadata standards and schemes, especially within the context of humanities computing and digital libraries, will also serve as complementary guides in the selection of initiatives and projects to explore and analyse. One example already mentioned is Lazinger's "Digital preservation and metadata: history, theory, practice"[99], which is a

[93] See e.g. Lazinger, 2001 for an extensive overview of issues in digital preservation and metadata and an annotated list of electronic social science data archives and cultural heritage digitization initiatives.
[94] See Geroimenko & Chen, 2002, Paulsson, 2003, Wiley, 2000; 2002,
[95] The drawback with XML is that everyone can develop their own standard or schema to employ in their own projects, with no necessary needs to take into account e.g. issues if interoperability with other domains.
[96] This opinion is expressed in McCarty, 2002 but also in DeRose, 1999 and Mylonas and Renear, 1999.
[97] http://www.tei-c.org/Applications/index.html where all the projects by can be sorted according alphabetically, by subject, corpus language or by date.
[98] http://www.humbul.ac.uk/
[99] Lazinger, 2001.

comprehensive guide to all sorts of digital resources in the humanities and social sciences worldwide. Another prominent example is "Digital resources for the Humanities" by Condron, Fraser and Sutherland[100] with a detailed overview of hundreds of digital resources for the humanities available on CD-rom and over the web, ranging from corpora, analysis tools and all sorts of virtual learning environment. Apart from the overviews and directories mentioned above there are also others overviews not as comprehensive and fit for my purpose and I will not elaborate them here.

However, the overviews and directories are not in a systematic way analysing across different initiatives or disciplinary residences, what the pitfalls have been or how the actual encoding and work have been done.

The corpus I will compile for analyse will thus consist of different types of textual data but belong to a single, although broad, semantic domain. Sampling will be organized around two tasks: i) defining the universe of relevant texts to be analyzed and ii) and a decision whether a full or partial coverage should be performed. My sampling decisions then will not only be directly related to the purposes of the text analysis but also affect the analysis results.

The methods and tools I intend to use for "drawing the maps" of the selected digital archives is text analysis by using a multivariate descriptive method and the topic map ISO standard.[101] Multivariate descriptive methods, such as correspondence analysis and different clustering techniques, provide means to describe and graphically visualise similarities among rows and column associations in contingency tables.[102] Topic maps provide means to build a semantic network above information resources, allowing visualization and navigation on a higher level of abstraction.[103]

What do I mean by textual data and what kind of information do I need to collect? [104] The general and short answer to this question may be any text which constitutes a relevant and necessary source material for answering the questions one is interested in. More specifically I will look for are all kinds of textual data that can be used for a structural (or concept) social text analysis such as, introduction texts about the project (aims, goals, etc.) guidelines, editorials, commentaries, articles and different kinds of reports etc.[105]

When the sampling is done I plan to use methods for analysing the content without prior categorisation (of the content to be analysed) in order to compare findings from the text analysis to the criterions already set. My aim is to try out relevant methods which can help to visualise the profiles of a series of text grouped into different criterions described below.

In this first step I will compile and analyse information available on the web *about the* initiatives and projects according to the three different strands I sketched on page 5.

---

[100] Condron, Fraser & Sutherland, 2001.

[101] The topic map was established as an ISO standard 1999. Topic Maps, International Standard ISO/IEC 13250:1999. International Organization for Standardisation (ISO), International Electrotechnical Commission (IEC). See also Gerionmenko & Chen, 2002. "Visualizing the Semantic Web: XML based internet and information visualization." for a useful overview of several different information analysis and visualisation techniques.

[102] Michael Greenacre, "Correspondance analysis and its interpretation" in Greenacre & Blasius, (Eds.) Correspondance Analysis in the Social Sciences, 1994:3-23. See also Ludovic Lebart, Exploring textual data, 1998:45-78.

[103] This kind of abstraction is similar to the aim with the Resource Description Framework (RDF) developed by the World Wide Web Consortium in 1999. See www.w3c.org/RDF

[104] See e.g. Ludovic Lebart, 1998 "Exploring textual data" for a comprehensive overview of relevant techniques.

[105] Of course, text is not the only thing which can be subjected to content analysis: images, films, music etc. are other kinds of communication which may be analysed for content; however, I will restrict myself to analyse textual data.

- *Content:* or what is considered to be the data and has been encoded? What has been collected? What is the rationale behind the choice of collected texts or items?
- *Tools and methods:* How have the resources been catalogued? What kinds of strategies, principles and interests have been informing the collection? What explicit principles have been guiding encoding and preparation?
- *Experimental practice*: In what kind of environment have the initiative been working? Is it a single person project or a larger collaborative effort?

What would be interesting is to reveal the morphology of collaboration - the structure and form of the different initiatives and how they relate to each other: e.g. stated aims, means and methods, on what premises and levels do they collaborate (I know they do to a certain extent), trans-national, inter-institutional, inter-disciplinary, how and where do they publish, who are the founders, associates and partners, who are the contributors and possibly why are the contributing.

The principle for analyse then, is to explore the possibilities to visualize a complex dataset of explicit and available information about resources using a new kind of method in combination with other techniques. This step will be an exploration both in how to find the nature of relations and both how to structure the already structured information. Hopefully, it might reveal salient relations and structures otherwise difficult to discern.

A method that might assist this type of analysis is topic maps. I shall not in any depth describe the method or the tools here, but rather focus on the elements topic map analysis might bring into my forthcoming survey so far: namely to incorporate the relational information between concepts.[106] That relational information will serve two purposes for this study; both as definition and both as their frequency of occurrence.

Similar to some corpus -based approaches in linguistics[107] topic maps are able to preserve the semantic structure of a text, and for that purpose it should be useful to examine how topics are related and not just how a certain topic, word or concept co-vary across texts within a corpus. The basic goal of my analysis is, given a set of texts and a set of topics, to analyse for each text whether a topic occur, as well as the relationships between the occurring topics.

The basic principles of topic maps are:

- topics: which is a construct corresponding to an expression of a real-world concept.
- occurrences: a topic may be linked to several information resources, such as web pages. The topic and their occurrences provide means to organise the information according to a specific concept. While the third component:
- associations: are describing the relationships between concepts.

Crucial in my topic map analysis is to determine what a topic[108] is and what a relationship is. A topic can be a word or a phrase and they can then be organized into types and hierarchically organized “downwards” in sub concepts or “upwards” into super-concepts. A relationship is a tie or a connection between concepts and it can be a single word or a clause, e.g. uses, are part of, belongs to etc.

By means of the above, using the topic map method to analyse texts, about the projects and initiatives, will results in a network of interrelated information with the topics being the

[106] Alexa, 1997.
[107] Cf. Biber, Conrad & Reppen (2002) *Corpus Lingvistics. Investigating Language Structure and Use.*
[108] I believe a topic in this sense can be understood similar to a concept.

nodes of the network and the relationships. The amount of information to be recorded for each association will be the analysts (i.e. my) choice. More specifically, one can choose to record that a relation do exists between two topics or, alternatively, one may decide to record the differences in all or some of the relationships. Naturally, the more information to be recorded, the more time-consuming and complicated the coding task becomes. Nevertheless, preserving a large amount of information clearly enables more detailed comparisons.

Detecting similarities and differences in the structure of the compiled texts may be assisted by frequency of co-occurrence information and estimations and differences of the distribution of topics and by exploring the nature of relationships among topics across the compiled texts (i.e. the corpus). A method that might be useful for detecting and analysing distribution and relations among topics is to use correspondence analysis. Correspondence analysis[109] is a systematic method for explorations of multidimensional data. The method can be used to investigate and describe magnitude and the nature of relations between row and column variables within cross tabulations or in binary tables.[110] An application of method was first introduced by Benzécri in 1973, who (according to Greenacre) use to say: "The model must follow the data not the other way around".[111] This is a technique primarily used to reveal features in the data instead of focusing on to reject or confirm hypotheses about structures and process from which data is generated. However, one need to make some assumptions about the data in order to be able to depict what is the data. This is a flexible method and can be used on many different kinds of data where a structure can be justified.

Combining the techniques of both methods and move from topic map analysis to correspondence analysis gives opportunities to alternate between statistically analyses of the coded texts and still remain close to the textual data and examining the occurring topics and with what specific relations are they connected to each other.

The topic map standard is intended to enhance navigation in complex data set. Although this standard allow organisation and representation of very complex structures, the basic concept of this model are "simple". One advantage is that the topic maps add semantics to existing resources without modifying them. A drawback is that the maps tend to be complex pretty soon and therefore difficult to navigate in.[112]

The drawback following this work would be that this kind overview might get out of date since the progress and the development is moving rapidly. Another drawback might be that I will found it hard to compare the texts compiled from different digital archives across different domains and contexts in a fruitful way.

A retrospective exploration

Since the one aim with my studies is to understand and enhance our knowledge about how researchers work, within their community but also across different boundaries and domains I intend to use some of the material already collected for the first study. The earlier mentioned TEI project (Text Encoding Initiative) will be subject to a retrospective exploration by the analysis of all sorts of records, meeting protocols, versions of talks and

[109] The brief description of correspondence analysis outlined in this paper refers to readings of Blasius & Greenacre (1998), Broady (1990) and Lebart (1998).
[110] A binary table is a table where one variable can have the value of either presence or absence not both - i.e. to have the value 0 or 1.
[111] Greenacre 1994:viii.
[112] Information about and topic maps specification can be found at: www.topicmaps.org and www.ontopia.net See Le Grand & Soto, 2001 for an introduction to the topic maps principles and applications. An extensive list of public available topic maps can be found at: http://www.topicmapping.com/registry.html

papers, articles in related journals. This exercise should also be useful to the related case study described below. The TEI Consortiums web page is a huge resource and archive being there waiting to be explored since all the documents from the first ten years of its existence, since 1988, have been assembled via servers and personal collections.

In addition all documents currently under production or produced after 1999 are available through the TEI website.[113] The archive consists of these documents; drafts of guidelines; committee documents, unnumbered reports; articles and presentations. Other records I intend to use in this study are related journals such as: Computers and the Humanities (Chum), Literary and Linguistic Computing (LLC). Conference proceedings and other relevant documents from the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL), and the Association for Literary and Linguistic Computing (ALLC) and Digital Resources for the Humanities (DRH).

This study will result in one article written together with Donald Broady.

Second step: Case studies of two research communities

Besides the overview of digital archives projects, it is important to explore how content mark-up and meta-data are used within some research domains. Therefore I intend to do case studies of two research communities with the purpose to gain some understanding in how teachers and students might make pedagogical use of resources that are already available on the web albeit mainly used by research specialists. I intend to choose communities from two different domains, the humanities computing community and the bioinformatics community focusing on the uses of their digital archives and affiliated applications of interchange languages for data. The main task will be to make comparative analyses of how existing digital archives and shared web resources actually are used within theses two communities. By contrasting these - at a first glance quite different - domains to each other I hope to gain comparative opportunities. This includes observing and exploring practical *uses* of actual content; uses and development of the tools and methods machinery at hand and how they collaborate and about what they collaborate.

The aim of the case studies is to analyze how members of communities develop their shared knowledge. How do research communities make use of digital archives and web repositories? How do researchers actually work with their tools and methods in the humanities communities on the one hand and the natural sciences on the other hand? In what way does knowledge-making in the humanities differ from knowledge-making in the sciences? Is there any kinship to be found?

Apart from the explicit records from documents and web pages that will be subject to analysis I will try to collect data (i.e. records) according to the questions I have posed but also regarding to the practical boundaries of this study. What possibilities do I have to collect the data? And what kind of data do I need to collect? In other words how do I deal with this kind of complex situation? If, I gain physical access (limited virtual access I already have in since some records are available on the web) to both of the communities then I envision that the plans at the moment are the following.

Collecting data in the case studies

I will start with the case study of the Oxford Text Archive (a part of the humanities computing community), by using observations and interviews. Here they have almost thirty years of collective experience of encoding and managing digital resources. I will conduct

[113] The TEI community have even prepared the data archive by indexing the documents to make it easier to find and use for anyone interested in writing the history of the TEI. See http://www.tei-c.org/History/index.html

observations during an initial three weeks period this autumn (2003) and hopefully I will have possibilities to do a follow-up study within the next two years period. The informants of the interviews will be selected within the humanities computing community (and not only the OTA) according to availability (pragmatic reasons) but I till try to interview those people, within the community, who are either consider having made major contributions to the community work or are representing contrasting views or perspectives (relative to each other and within the community). These people are not very difficult to discern, in fact by reading the guidelines and articles in the journals and studying the conferences proceedings I have been able to depict possible candidates for my forthcoming interviews. One important aspect and a difficulty I may encounter in gaining access to this community is that they do not know who I am - a totally new person who wishes to be a member of this community and at the same time would like to investigate what they are doing.

Here there are some issues which have to be considered. Firstly; how do I gain access to the intended informants and "trust" in this community and how to address the informants accordingly? Secondly; how do I handle the fact of investigating the same community that I want to be a part of? This might be more difficult than I initially thought of and it has to be dealt with in a sensitive way. Bourdieu begins *Homo Academicus* with a discussion of the particular problems inherent in studying ones own environment (I am not comparing this situation to what he is describing though). He discusses the epistemological challenges involved in breaking with everyday experience and reconstructing the knowledge obtained in the first break.[114] He also discusses the self-reflection of this process, noting that researchers who study the same realm where s/he self operates can use the results from the research – and so to speak reinvest the results - as tools for self-reflection of the limits of the research and to understand what s/he is motivated to see and not to see.[115]

The other case study is planned to take part within the bioinformatics community in Sweden. This community I hope to gain access to through my earlier work in and collaboration with the Bioinformatics project.[116]This study might be on accomplished during a longer time period, but the observations will not be done any more than approximately three to four weeks altogether. Finding suitable and be introduced to possible informants for the interviews will be due to my earlier contacts with leading scientists in this field. I am not sure if there will be a problem to gain access to one or two physical places where I can do the observations.

Considerations concerning both case studies

The problem I believe will rather be how to collect records of what they actually are doing when they are working with their digital archives. Here I envision that I will ask them to describe out loud what they are doing and then collect different representations of their work.

During the interviews I will ask the researchers/informants to describe their recent work and to discuss the information processes involved in a specific and recent activity/task (such as doing an analysis, adding a record or preparing teaching[117]) related to the person being interviewed and related to the actual use of digital archive(s): e.g. what was used; how was it used; and how the interviewee worked through the activity including any the

[114] Bourdieu, 1996a: 35.
[115] ibid. p.48
[116] This project was formally named the Life Science Project and was a three year project (1999-2001) within the Wallenberg Global Learning Network funded by the Knut and Alice Wallenberg Foundation. See Andersson, Langerth Zetterman & Strömdahl, 2001 for a description of work undertaken in this project.
[117] Associated courses within these communities might also serve as a part of the empirical material.

final stages. I think the explanatory-oriented approach might be useful in helping informants to think about and describe their work in specific terms.[118]

Hopefully I will be able to do tape recordings during the interviews, which require cooperation from the informants, but it should be possible since researchers who are "confident of the positive knowledge in their fields" often have an open attitude and therefore have no wishes to hide that knowledge.[119] On the other hand this method is very labour intensive since it requires someone (i.e. myself) to transcribe the tapes and to learn a transcription method.

In addition to records from observations and interviews I will use the already collected data from the other studies in the first case (humanities computing/TEI/OTA) and in the other case (bioinformatics) I will collect records addressing uses of their digital archives, such as articles, presentations, syllabus, internal notes etc. This kind of records might be somewhat fragmentary but better than no records at all.[120]

When analysing uses of digital archives I intend to use a procedure where I add descriptions of each of the collected items.[121] When compiling and combining selected items, e.g. different kinds of texts and protocols from observations, with the descriptions of usage of a particular item in a table, it might be a useful mechanism for learning how researchers are using specific sources as well as to find out any attributes or relative significance that may be attached to the uses and/or items.[122] Moreover, using this procedure might enable a possible documentation on how arguments are constructed in relation to their use.[123]

The case studies will result in two articles, were the first one probably focus on methodological and data collection issues and the second one focus on the data analysis and the results.

Third step: Application of tools and methods for content design

Besides, the above mentioned mapping of actual practices within research I also wish to explore and to some extent promote the application of a few tools. This will be accomplished by setting up an experimental environment in which a user study focused on testing of some tools aimed for manipulation, encoding and provision of digital content on the web and the development of an experimental (and very modest) digital archive. Here I will gain from previous experiences from earlier mark-up activities and work within Digital Literature[124] and from the extensive body of experiences and good practices I have hopefully will find through the initial overview.

An experimental content archive will be built on a few resources/texts, marked up according to the TEIXLITE DTD (the Text Encoding Initiative Document Type Definition). The archive will initially consist of different kinds of texts, such as a research report on Social Science Classics[125] and selected chapter(s) from a book in the history of

[118] Brockman et al. and Knorr Cetina (1983 & 1999) used this kind of approach successfully in their case studies.
[119] Knorr Cetina, 1999:21.
[120] See ibid. pp. 21-23 for a discussion on collecting records.
[121] What would be considered as an item is yet to be decided. For instance: should a new version of text in the digital archive be treated as a new item or as an(other) instance of the same item?
[122] For example each of the items collected will be equipped with a description (metadata) such as type, format, name, date, subject/content, location, role.
[123] A similar procedure has been used in Brockman et al. (2001)
[124] See e.g. Broady, 1996; 1997; 2001 and Juliusson, 1997.
[125] Broady, 1998. "Läsestycken för samhällsvetare i urval och översättning av Donald Broady". 5 upplagan.

philosophy[126], and biographies created within the research project: Formation for the public sphere (FFO). Eventually this archive also might be including the digital editions of August Strindberg's work prepared and encoded in SGML in 1996-1998. [127]

The subject to be addressed here will foremost be focused on issues in the design and encoding of the collective biography with the TEI DTD. The encoding is done on newly produced and changing research material, i.e. all the biographies compiled to a large collective biography instead of already existing text. The design of the textual database will also be subject to an analysis and a comparison with other similar projects and initiatives.

This study is supposed to result in one or possibly two articles. At least on of the articles will be written together with Donald Broady.

Formation for the public sphere –exploring a collective biography

One study will be about the experimentation with and an exploration of a specific kind of digital archive: a collective biography in the project "Formation for the public sphere" were detailed biographies on bourgeois women, around the turn of the century 1900 in Sweden are produced and compiled.[128] In this study I will focus on methodological issues during designing and encoding of the collective biography and historical database. The material consists of data collected with a prosopographical method and the projects aim is to study the history and structure of this field rather than the individual woman.

The scholars in the project are working physically dispersed at different universities and departments. Therefore they need a stable web based repository and working environment. The research group has chosen to use the web based platform (BSCW) for handling all the documents produced. Therefore I will also to some extent studying the work process using this platform aimed for collaboration and sharing of common documents. A key issue here is possibilities for versioning handling of the collectively produced documents. Besides, it provides a private area for the collective since the project participants do not want to publish work in progress on the web but nevertheless need to be able to reach the documents from any available computer.

The period around the turn of the century 1900 was of crucial significance for the women's way from the private to the public sphere. In the research program "Formation for the public sphere" the aim is to study the impact and importance of meeting places and social networks for women in entering the public sphere in Sweden. The focus in the project is on the women's contributions in philanthropy, health-care, culture, and education during the period of 1880 to 1920. The goal is to create a historical database based on a collective biography of the bourgeois women, encoded according to the TEIXLITE DTD for transformation into a database enabling further export to different software used for statistical analysis.

A method called prosopography[129], is used to collect detailed information on all women (about 100) aimed for analysing the women's careers, strategies and formations. The

[126] Andersson, J. (1998) Filosofisk tanke. 2:a upplagan.
[127] See Broady, 1996.
[128] Donald Broady, Uppsala University; Boel Englund, Uppsala Univeristy; Ingrid Heyman, Uppsala University; Agneta Linné, Stockholm Institute of Education; Kerstin Skog-Östlin,University of Örebro; Eva Trotzig, Stockholm Univeristy Library; Annika Ullman, Stockholm Institute of Education "Formation for the Public Sphere. A Collective Biography of Stockholm Women 1880—1920". A research program, 2000 Available in Swedish: http://www.skeptron.ilu.uu.se/broady/sec/p-ffo-00.htm
[129] See Broady, 2002 for definition of prosopography. The study on the Parisian academic field is an example of prosopograhy. Pierre Bourdieu: *Homo academicus*. Paris: Minuit, 1984, English translation: *Homo academicus* Cambridge: Polity Press, 1988.

prosopographical method is developed for study of individuals belonging to the same field, i.e. a certain social group. It is based on a comprehensive collection of data were the purpose is to collect information in the same categories on all individuals included in the sample. This allows to draw a "map" of the distribution of characteristics of the individuals by using the correspondence analyse method and then allowing analysis of strategies and acting on different social fields. The main object is not the analyse of the individuals (women) per se but rather the history and structure of the field itself.[130]

Thus, detailed information is collected on each woman on standpoints in matters important to the field and other characteristics relevant to the analysis; e.g. social origin, educational capital and way though the formal educational system, marital status, symbolic, economic and social capital and also her social and cultural practices. Examples on the kind of information being surveyed are: where she grew up and where she lived, who here parents where, whom they were associated with, her friends, her occupation, which her published writings if any, which associations she was engaged and affiliated with, etc.

When approximately a third of the biographies had been written they were evaluated with the purpose of making a decision on the minimal set of information categories needed for the database. Altogether about fifty categories and variables were decided to be included in the minimal dataset, for each woman, to be encoded in the first step. The actual encoding, which includes some editorial work, is done by one and the same person. In the first step approximately between fifty to sixty biographies will be encoded with the minimal dataset. The key issue here is to encode the material in such a way that all the categories will be covered in a purposive way allowing for a possible extension of both more biographies and encoding added to the minimal dataset.

In order to accomplish the goal of analysing the material and building the database, work has been undertaken in several different steps. Firstly, the scholars in this project have written the biographies using a structured schema, which was written in plain text allowing interoperability and transformation from the scholar's choice of word processor (usually MSword). The structured schema is used to ensure that all necessary information about the women is accounted for and to ensure congruency in the way the biographies are written by the different scholars. The schema is also of great help for the encoding since the information is fairly consistent and organised in a similar way.

Secondly, all individual biographies are compiled and encoded, including a certain amount of editorial work, by another person (me) not writing the actual biography. The collective biography will be available on the web for research and educational purposes.

Thirdly, the requirement specification of the database has been decided by the scholarly collective in the project and the encoder's task is to prepare and encode the material accordingly and to map data required for statistical analysis into a database.

The task then will be to map the encoded data into a relational database for further interchange and transport to the statistical software used for the correspondence analysis.[131] The encoding and the database design process is guided by recurrent discussions, among scholars and the encoder, about e.g. editorial and content issues regarding the individual biographies produced as well as the collective biography.

In this study the reciprocal dependency between the content design and the methods used in the research will be investigated. to consider the extent on TEI, designed to describe and

[130] According to the characterise the prosopographic studies undertaken by Bourdieu and his followers Centre de sociologie européenne, École des Hautes Études en Sciences Sociales, Paris

[131] Of use here might be e.g. Simons, 1999. "Using Architectural Forms to Map TEI Data into an Object-Oriented Database."

encode existing texts, is also suitable for the design of new kinds of material. Other issues of interest might be to reflect on the ways in which the technical methods can aid researchers in posing new questions or new lines of enquiry.

Second article

For the second study/article in this step I have no specific plans at the moment. I will probably focusing on exploring uses of tools and methods (semi-automatic and manual techniques) and related issues. For instance to consider if existing encoding guideline such as the TEI, designed to describe and encode existing texts, is also suitable for the authoring other kinds of material, such articles, web sites or academic papers. Investigating the availability of tools and methods for a larger community of researcher and teachers might also be of interest.

A preliminary project time plan

Year	Semester	Research activity	Article
1	Spring 2003	Experimental content archive (d1) & overview (c1)	(drafts)
2	Fall 2003	Overview (c2) & Case study (a1) (observation, interviews)	1
	Spring 2004	Experimental content archive (d2) & overview (c3)	2 & 3
3	Fall 2004	Case studies (a2, b1)	4
	Spring 2005	Case studies (a3, b2 & b3)	5 & 6
4	Fall 2005	Experimental content archive (d3)	7
	Spring 2006	Finish up & dissertation	

Activities:

a) humanities computing case study
b) bioinformatics case study
c) overview
d) experimental archive

Articles:

1. The first overview of digital archives (methods with some initial results from the overview)
2. The first article on the experimental archive with a analyse of tools and methods used related to the encoding of biographies and the design of a textual database of that source.
3. The second overview – the map of the territory of digital archives using the TEI DTD (and similar full text archives and initiatives).
4. The first article on case study focusing on unfolding the research communities and on methodology
5. The third overview – a retrospective study of the TEI community
6. The second article in the case study presenting the results including the comparison between the two communities
7. The second article on the experimental archive addressing experiences in content design

References

[to be completed]

Andersson, S., Langerth Zetterman, M., & Strömdahl, H., (2001). Theory-anchored evaluation applied to a CSCL intense course in bioinformatics. *In proceedings of The First European Conference on Computer-Supported Collaborative Learning, Euro-CSCL, Maastricht, March 22 - 24, 2001*.

Alexa, M. (1997). *Computer assisted text-analysis methodology in the social sciences.* Mannheim: ZUMA Arbeitsbericht. 97/07.

Alexander, S & McKenizie, J. (1998). *An Evaluation of Information Technology Projects for University Learning.* Canberra: Australian Government Publishing Service. Executive summary available at: http://www.dest.gov.au/archive/cutsd/publications/exsummary.html [2003-03-01]

Blanton, W., Moorman, G. & Trathen, W. (1998). Telecommunications and Teacher Education. In D. Pearson & A. Nejad-Iran (Eds.). *Review of Research in Education, 23,* 235-276.

Broady, D. (1996). *Digital Critical Editions. The case of the Swedish National Edition of August Strindberg's collected works*. Paper read at the conference DRH96 (Digital Resources for the Humanities), Oxford University, July 1-3 1996.

Broady, D (1997). *Content design. Methods and tools for the creation of portable hypermedia archives. Notes for a proposed CID project*. Version 2, 1997-10-04. Available at: http://www.skeptron.ilu.uu.se/broady/dl/p-broady-cd-971004.htm [2003-03-01]

Broady, D. (2001) "Digitala arkiv och portföljer", pp. 11-16 i *IT i skolan - mirakelmedicin eller sockerpiller?* Rapport 45/2001. Stockholm: IT-kommissionen, 2001. Available at: http://www.skeptron.ilu.uu.se/broady/dl/p-broady-digark-01.htm [2003-04-02]

Broady, D. & Haitto, H. (1996). *Internet and the humanities: the promises of Integrated Open Hypermedia*. Paper read at the conference "Contemporary computer and network technologies", Moskva, 17-18 Jan. 1996. (Report IPLab-106, jan 1996, and Report CID-1, 1997). In *Proceedings of Contemporary computer and network technologies*, Children's Computer Club, Moscow, January 1996, digital publication.

Brockman, W.S., Neumann, L., Plamer, C. & Tidline, T. (2001). *Scholarly work in the humanities and the evolving information environment.* Washington, D.C.:Digital Library Federation. Council on Library and Information Resources.

Bourdieu, P., Chamboredon, J-C., & Passeron, J-C. (1991). *The Craft of Sociology. Epistemological Preliminaries.* (Ed. Beate Krais). Berlin/New York: Walter de Gruyter. (Transl. of *Le métier de sociologue*, 2nd edition 1973)

Bourdieu, P. (1996a) *Homo academicus*. Stockholm/Stehag: Brutus Östlings Bokförlag Symposion.

Bourdieu, P. (1996b). The state nobility. Elite Schools in the Field of Power. Cambridge, UK: Polity Press.

Burnard, L. (2001). *From two cultures to digital culture: the rise of the digital demotic.* Avaliable at: http://users.ox.ac.uk/~lou/wip/twocults.html [2003-04-14]

Chambers, E. (2000). Computers in Humanities Teaching and Research. *Computers and the Humanities, 34* (3), 245-277.

Condron, F, Fraser, M. & Sutherland, S. (2001). *Oxford University Computing Service Guide to Digital Resources for the Humanities.* Morgantown: West Virginia University Press.

Cook, S. & Brown, J. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science, 10* (4), 381-400.

DeRose (1999). XML and The TEI. In *Computers and the Humanities, 33*. 11-30.

Fils, D., Taber, R, Takle, S. & Sorensen, E (2000). *Distributed Collaborative Learning across Disciplines and National Borders: Structuring through Virtual Portfolios.* Available at: http://www.hum.auc.dk/~vipol/papers/elsebeth/portfolio.htm [2002-12-11]

Flanders, J. (2002). Learning, reading, and the Problem of Scale: Using Woman Writers Online. In *Pedagogy: Critical Approaches to Teaching Literature, Language, Composition, and Culture*.(2)1, 49-59.

Frazer, M. (2000). From Concordances to Subject Portals: Supporting the Text-Centred Humanities Community. *Computers and the Humanities, 34* (3), 265-278.

Fujimura, J. (1992). Crafting Science: Standardised Packages, Boundary Objects, and "Translation". In A. Pickering (Ed.). *Science as Practice and Culture.* Chicago: The University of Chicago Press.

Gerionmenko, V. & Chen, C. (2002). Visualizing the Semantic Web: XML based internet and information visualization. London: Springer.

Hacking, I. (1992). The Self-Vindication of the Laboratory Sciences. In A. Pickering (Ed.). *Science as Practice and Culture.* Chicago: The University of Chicago Press.

Hacking, I. (1999). *The Social Construction of What?* Harvard University Press: Cambridge MA.

Hockey, S. (2000). *Electronic Texts in the Humanities.* New York: Oxford University Press Inc.

Illich, I. (1970). Schooling: the Ritual of Progress. The *New York Review of Books.* December 3, 1970.

Illich, I. (1971a). *Deschooling Society*. Originally published: New York: Harper & Row; London: Calder & Boyars. Avaliable at: http://philosophy.la.psu.edu/illich/deschool/intro.html [2003-04-16]

Illich, I. (1971b) A Special Supplement: Education Without School: How It Can Be Done. The *New York Review of Books.* January 7, 1971.

Illich, I. (1991). *Text and University. On the idea and history of a unique institution.* Translation by Lee Hoinacki of the keynote address delivered at the Bremen Rathaus, September 23, 1991, on the occasion of the twentieth anniversary of the founding of theUniversity of Bremen. Available at: http://alf.zfn.uni-bremen.de/~pudel/index.html [2003-04-02]

Iivonen, M., White, M. D. (2001) The choice of initial web search strategies: A comparison between Finnish and American searchers, *Journal of Documentation* (57)4, 465-491.

Juliusson, J. (1997). *SGML-märkning av medeltida jordeboksmaterial*. Master thesis TRITA-NA-D9705, CID-14, Nada, May 1997.

Knorr Cetina, K. (1999). *Epistemic Cultures; How the Sciences Make Knowledge*. Cambridge, Massachusetts; Harvard UP.

Knorr Cetina, K.& Mulkay, M (1983). *Science observed.* London: Sage.

McGann, J. (1998). Textual Scholarship, Textual Theory, and the Uses of Electronic Tools: A Brief Report on Current Undertakings. In *Victorian Studies*. Summer. 609 – 620.

Mylonas, & Renear, A. (1999). The Text Encoding Inititive at 10: Not just an Interchange Format Anymore – But a New research Community. In *Computers and the Humanities, 33.* 1-9.

Langerth Zetterman, M. (2001). IT-stöd i distansutbildning med fokus på lärande - Nya förutsättningar och konventionella lösningar. *Pedagogisk forskning i Uppsala 141.* Pedagogiska institutionen: Uppsala universitet

Langerth Zetterman, M. & Lindblad, S. (2001). Learning about e-learning. A starter about Internet discourses and borderless education. *I NFPF:s 29:e kongress Stockholm. Pedagogikens mångfald. Lärande innanför och utanför institutionerna. 15-18 mars, 2001.*

Laurillard, D. (1993). Rethinking University Teaching. A Framework for the Effective Use of Educational Technology. London: Routledge.

Laurillard, D. (2000). New technologies, Students and the Curriculum. The impact of Communications and Information Technology on Higher Education. In P. Scott (Ed.) *Higher Education Re-formed.* London & New York: Farmer Press.

Lazinger, S. (2001). Digital preservation and metadata: history, theory, practice. Englewood, CO: Libraries Unlimited

Le Grand, B. & Soto, M. (2001). Topic Maps Visualization. In V. Gerionmenko & C. Chen (Eds.). *Visualizing the Semantic Web: XML based internet and information visualization.* London: Springer

McCarty, W. (1998). *What is humanities computing? Toward a definition of the field.* Article presented at Liverpool, 20 February 1998. Reed College (Portland, Oregon, U.S.) and Stanford University (Palo Alto, California, U.S.), March 1998. Würzburg (Germany), July 1998. Available at http://www.kcl.ac.uk/humanities/cch/wlm/essays/what/ [2003-04-19]

McCarty, W. (2002). Humanities Computing: Essential Problems, Experimental Practice. *Literary and Linguistic Computing, (17)* 1, pp 103-125.

McGann, J. (1998). Textual Scholarship, Textual Theory, and the Uses of Electronic Tools: A Brief Report on Current Undertakings. *Victorian Studies* 41 (Summer).

McNair, V. & Galanouli, D. (2002). Information and Communication Technology in Teacher Education: can reflective portfolio enhance reflective practice? *Journal of Information technology for Teacher Education 2* (11) pp.181-196.

Meyer, K. & Tusin, F. (1999). Pre-service Teachers' Perceptions of Portfolios: process versus product. *Journal of Teacher Education, 7,* pp.131-139.

Paulsson, F. (2002). Standardized Content Archive Management – SCAM: Storing and distributing learning objects and learning components. Avaliable at: http://www.skolverket.se/skolnet/texter/scam_eng.pdf [2003-03-01].

Paulsson, F. (2003). Komponent baserade lärmiljöer och lärteknologi standarder. Utkast/internt arbetsmaterial.

Schön, D. A. (1988). Designing: rules, types and worlds. *Design Studies, 9* (3), 181-190.

Simons, G. (1999). Using Architectural Forms to Map TEI Data into an Object-Oriented Database. In In *Computers and the Humanities, 33.* 85-101.

Sjunnesson, J. (2001). *Digital Learning Portfolios: Inventory and Proposal for Swedish Teacher Education*. Uppsala University: ILU. (BA-thesis)
Available at: http://www.skeptron.ilu.uu.se/broady/dl/p-sjunnesson-digitallearningportfolios-0201.pdf [2003-04-16]

Sjunnesson, J. (2003). Metadata for learning objects on the web: overview, prospects and test. Master thesis. (2nd draft, March 2003).
Available at: www.skeptron.ilu.uu.se/broady/dl/p-sjunnesson-metadata-030311.doc [2003-03-19]

Spaeth, D. & Cameron, S. (2000). Computers and Resource-Based History Teaching: A UK Perspective. *Computers and the Humanities 34* (3) 325-343.

Tolsby, H. (in press) Digital Portfolios: a Tool for Learning, Self-Reflection, Sharing and Collaboration. Available at: http://www.hum.auc.dk/~hakont/ [2002-12-11]

Sutherland, K. (1997). Electronic Text. Investigations in Method and Theory. Oxford: Clarendon Press.

Wiley, D. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor and a taxonomy. In D. Wiley (Ed.) The instructional Use of Learning Objects. Available at: http://reusability.org/read/chapters/wiley.doc [2003-04-02]