Digitizing Intangible Cultural Heritage

Uneson, Marcus; Wittenburg, Peter

2004

Marcus Uneson
Department of Linguistics, Lund University
Lund, Sweden

Peter Wittenburg
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Report on the course

Digitizing Intangible Cultural Heritage

Vilnius, Lithuania, March 15-20, 2004

Phase III of UNESCO project

Establishment of a National Inventory and Electronic Database of Lithuanian Intangible Cultural Heritage

Version 1 -- FINAL
Oct 8, 2004

Abstract

As part of the UNESCO project "Establishment of a National Inventory and Electronic Database of Lithuanian Intangible Cultural Heritage" the authors, representing the EU-funded project "European Cultural Heritage Online" (ECHO) were invited to give a course in digital archiving called "Digitizing Intangible Cultural Heritage" in Vilnius, Lithuania, March 15 to 20, 2004. The present report summarizes very briefly the sessions given. Thereafter, the analyses of the state of the digitization work of the participating institutes and recommendations for the future are given in a dedicated, stand-alone section.

Background

As part of Phase III of the project "Establishment of a National Inventory and Electronic Database of Lithuanian Intangible Cultural Heritage", founded by UNESCO[1], the authors offered a course in digital archiving called "Digitizing Intangible Cultural Heritage" in Vilnius, Lithuania, March 15 to 20, 2004.

The contacts were established via Lund University as being a member of the current EU-funded project ECHO (European Cultural Heritage Online, 18 months, Dec 2002 -- May 2004)[2]. The main goal of ECHO is the establishment of a European infrastructure fostering the transfer of cultural heritage to the internet, permitting free access to fully interoperable, standards-compliant corpora of primary cultural heritage documents, as well as tools to exploit these documents. Both authors have worked with ECHO since its beginning, Mr. Wittenburg at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, and Mr. Uneson at Lund University, Sweden. In the preparations and afterwork, Mr. Uneson acted as main coordinator on the ECHO side, whereas Mr. Wittenburg carried out at least as large a share of the actual teaching.

The course was held at the Lithuanian Folk Culture Centre[3] (LFCC), B.Radvilaites g. 8, centrally located in Vilnius. Local coordinators at LFCC were Vida Satkauskiene and Dalia Usinaviciene, with Sigitas Jonkus as technical expert. Coordinator on the UNESCO side was Anahit Minasyan, Division of Cultural Heritage/Intangible Heritage Section[4].

The course attracted some 30 participants, although not all were able to attend all sessions. They represented various institutes dealing with Lithuanian cultural heritage, such as the LFCC[5], the Institute of Mathematics and Informatics[6] (Unesco Chair of Informatics for the Humanities), the Institute of Lithuanian Literature and Folklore[7], the Lithuanian Academy of Music[8], the Lithuanian Institute of History[9], the Library of the Lithuanian Academy of Sciences[10], and the Institute of the Lithuanian Language[11]. A list of participants is given in Appendix A. Judging from the evaluations we received, the course was generally much appreciated[12]. The authors certainly were much impressed by the enthusiasm and creativity shown by the participants, and we learned a lot about Lithuanian culture.

This report is meant to give a comprehensive view of the activities and, in particular, to provide recommendations for further work. The section on Analysis and Recommendations is phrased to be reasonably stand-alone; it could thus be independently circulated as a separate document, should this be more convenient. Links and pointers to further information are given under "Links" section below. General course information, slides, links etc may be found on the web site of the course[13].

Programme

Programme of visit

We arrived on Sunday, March 14 -- in Mr. Wittenburg's case, just in time to get a first glance of Lithuanian folklore at an event at LFCC. Monday was reserved for informal talks at LFCC with

the local coordinators (Ms. Satkauskiene, Ms. Usinaviciene, Mr. Jonkus) about the planning of the rest of the week, as well as a tour through LFCC, demonstrations and examples of the kind of material LFCC houses and of the current digitization practice. These talks and demonstrations led to certain revisions of the programme (see below).

The course programme for Tuesday to Friday is given below and commented in some detail later. Outside the program, we might mention that we were invited to lunch with the representatives of the Lithuanian National Commission for UNESCO, Ms. Asta Dirmaite and Ms. Milda Paskauskaite, on Thursday, giving us opportunity to briefly explain the ideas behind metadata interoperability and the ECHO project. We were kindly invited to visit the offices of their organization; however, the rather dense program unfortunately did not allow this.

Digitization activities aside, our Lithuanian hosts, Ms. Satkauskiene, Ms. Usinaviciene, and Mr. Jonkus, continuously showed their hospitality. We were invited to several evening concerts and performances, and even honoured with a small ceremony at course end, including an improvised concert. Our hosts were also kind enough to take us on an excursion to the beautiful Trakai¤ area on Saturday before we finally left on Sunday, March 21.

Course programme

Since a course of this kind was a new experience to both authors, we had tried to set up the programme somewhat interactively, by dedicating a web site[14] to the course, publish a draft programme, and ask for feedback. However, it soon became clear that the programme thus arrived at would put unnecessarily much emphasis on digitization itself. During the informal planning talks at LFCC on Monday, we found this phase to be a mostly solved problem. We concluded that the participants' time could be better invested by focusing more on the questions that follows digitization: metadata and metadata interoperability, annotations, presentation vs representation, container models, etc. The program was revised accordingly. This meant some reshuffling and at times a bit of improvisation, especially for the slides, but we do believe that the revised program was better suited to meet the needs of the participants.

It was also clear that the planned hands-on training with, e.g., metadata editors had to be omitted for practical reasons -- too many people on not enough available computers. It is anyway the experience of Mr. Uneson, who would have been the instructor of this part, that such practical training not always is very efficient in larger groups -- especially, as here, when the degree of computer literacy varies --, and that a commented demonstration by the teacher may provide a sufficient foundation for individual training in the tempo and environment of the participant's choice.

We tried to adapt as best we could to the wishes of the participants; this meant several later revisions, including reserving Saturday exclusively for any technical questions that had not been answered so far (mostly of concern for system administrators and similar). This arrangement permitted most of the participants to finish the course on Friday evening. The program we finally arrived at, besides the more informal talks on Monday and Saturday, is given in Table 1 (although little attention was paid to the details). The main topics are summarized in the following section.

Table 1. Course program

Time	Tuesday, March 16	Wednesday, March 17	Thursday, March 18	Friday, March 19
Theme	**Planning and building a digital archive**	**Metadata & interoperability**	**Archiving principles Workflow Distribution**	**Content Project administr. Project management Access issues Technical questions**
Morning session I 9.00 - 10.30	Welcome short presentationsm practical matters, course overview Digitizing and its challenges MPI & ECHO NINCH principles MPI archive Delaman & Grid	Metadata Introduction DC, IMDI	Q & A on metadata & interoperability	Content: annotation and analysis tools (Praat, ELAN, Annotea, Transcriber)
Morning session II 11.00 - 12.30	cont.	cont.	Archiving principles Workflow File formats	Archive management Access rights (legal, ethical) Access management
Afternoon session I 13.30 - 15.00	Planning a digitization project Planning a digital archive Talk and discussion -- your needs	Metadata sets Interoperability ECHO interoperability solution	cont.	Technical archiving solutions (cont) Q & A
Afternoon session II 15.30 - 17.00	cont.	cont.	cont.	cont. + Wrapping-up discussion, any questions

Session summaries

Tuesday. Planning and building a digital archive

Mr. Uneson bid everyone welcome to the course and all participants were asked to introduce themselves and their backgrounds briefly. After short presentations of their respective home institutes (Max Planck Institute for Psycholinguistics[15], Nijmegen, and the Department of Linguistics, Lund University[16]) and of the ECHO project which they represent, Mr. Wittenburg and Mr. Uneson gave an introduction to digitization of cultural heritage, and an overview of the course.

Mr. Wittenburg started by showing examples from the linguistic area of the urgency of immediate recording and documentation. According to some estimations, he said, 90% of the languages spoken today will be dead in a hundred years. He went on to the general rationales for digitization, concentrating on its potential of saving valuable material currently housed on deteriorating media, but also its being a prerequisite for many operations necessary in modern research methods and for efficient data storage and distribution. The many ongoing and concluded digitizing projects has created huge digital collections, and still we only see the beginning. There are also traps, however; not always have important questions such as "what?", "how", "for whom?", "to what cost", "to what quality?" etc been given sufficient attention. As Mr. Wittenburg put it, "brainless digitization creates a digital data cemetery". He mentioned the "local trap", meaning that the creators focus their intentions on short-term goals, such as presentations in a predefined format; or on certain tools, without which the resources cannot be exploited. In both cases later ingestion of the resources in a larger context may be very expensive or even impossible, and the data may never be used outside the small group of creators.

Another trap is the "big pot dumping", where too much resources are put into producing the digital resources themselves and too little on tasks needed to make them useful in the first place, such as management, structuring, storage, distribution, annotation, metadescriptions, etc. Big pot dumping may mean that existing resources are never exploited, because they need much more preparation to be useful, or even because they cannot be found at all.

Mr. Uneson, starting with the World Bank's very wide definition of Cultural Heritage, then described the pillars of the ECHO project and its aim to create and explore a new IT-based cooperative research infrastructure. He gave a few real-world examples of ECHO material, such as digitized texts in classic languages with automatic morphology analysis[17]; digitized maps with tools for cooperative image annotation[18]; and examples of reconstructed historical buildings. He then summarized the conclusions of the ECHO state-of-the-art report[19]: a) there is currently a general fragmentation of digital and digitized material, meaning an enormous waste of resources; b) well-designed, interoperable metadata sets with good tools may do much to overcome this waste; c) tools should preferably be developed in dialogue with their future users.

The achievements of the ECHO project are impressing; however, Mr. Uneson stressed that in any long-term perspective, we should not think in terms of projects, with a certain date of initiation and termination; such planning may have us falling straight into Mr. Wittenburg's "local trap". Rather, he said, we should plan at an early stage for data reuse and sharing in user-friendly, searchable, browsable, open standards-based, interoperable domains. He concluded by informing the participants that the soon-to-open Centre for Languages and Literature at Lund University[20] may offer to host cultural heritage data for institutes that lack resources themselves.

The speakers then offered two real-life projects as examples. Mr. Wittenburg described the extensive DoBeS (Dokumentation der Bedrohten Sprachen -- Documentation of Endangered Languages) project, funded by the Volkswagen foundation and documenting not only language, but also oral folklore, music, etc. He also showed many images and played audio and video recordings collected in this project. Mr. Uneson mentioned the St. Laurentius project of Lund University, a rather typical digitization project of the around 70 medieval manuscripts at Lund University Library.

The speakers then went on to introduce project and archive planning. Mr. Uneson introduced the Golden Rule of Digitization: "Aim for broadest use of the material, today and in the future, perhaps including unimagined audiences and applications; do so by applying open standards and community-shared good practice". He took this rule as a point of departure to arrive at the six core principles outlined in the important work of the National Initiative for a Networked Cultural Heritage (NINCH) and documented in the useful Ninch guide[21]:

1. Optimize interoperability of materials
2. Enable broadest use
3. Address the need for the preservation of original materials
4. Indicate strategy for life-cycle management of digital resources
5. Investigate and declare intellectual property rights and ownership
6. Articulate intent and declare methodology

He reminded the audience about the importance of asking the Journalist's Questions at the start of a project: what is the aim? who will benefit? what material should be selected? how should it be done? what standards, what metadata, what equipment, what workflow? who should do it? and when? and for what money? etc.

Mr. Wittenburg then went on by putting such a project in a larger and more long-term perspective, as he raised several important issues on life-time aspects in digital archiving. He made clear that although not all long-term aspects of a digital archive are entirely under the control of the archivist (e.g., political decisions), those that are should be duly considered. As three cases in point, one might think how to guarantee the physical survival of data, the interpretability of the data, and the long-term management of the data in terms of social and political uncertainties--who knows how long the archive-holding institutes will exist?

For physical survival and the uncertain long-term fates of archives, Mr. Wittenburg compared expected life times of some relevant entities, such as hard disks, CD-ROMs, the Max Planck Society, and the nations housing them. In a long-term perspective, he said, it is clear that automatized copying between mass storage devices, and continuous migration between them, are necessary to compensate for deteriorating media. Likewise, distributed, international copying may remedy data loss due to social or political developments. As Mr. Wittenburg pointed out, data within the Cultural Heritage and Humanties communities tend to be rather small compared to that of, say, physicists; for large computing centres providing archiving and backup services to such communites, the extra cost of also including data from the humanities is rather low. He described the current fully automatized setup at MPI-PL, where two copies are stored on local servers, one in an adjacent building, one in Munich and one in Göttingen. On a European level, he mentioned the emerging GRID automatic, international copying initiative. As for interpretability, he stressed the importance of documentation, adherence to (open) standards, and general awareness of the problems. Both authors gave some examples from the unfortunately rather rich catalogue of bad examples from the history of digitization.

Mr. Wittenburg further pointed out that there often is a certain tension between the goals of researchers, aiming at rapid data exploitation, and those of archivists, aiming at long-term data preservation, and concluded that focus for a long-term commitment must be representation rather than presentation. He reminded the audience that tools (that is, specific-purpose programs to handle the data) generally are rather short-lived, typically 5-10 years -- no long-term archive should be dependent on a specific tool, especially not on a proprietary one. Having said this, however, Mr. Wittenburg also pointed out that good, standards-adhering, non-propriatory tools may be extremely helpful and indeed often absolutely necessary: they may determine efficiency, consistency, and standards compliance, they may help in management, etc. He concluded by expressing his intention to return to several of the themes mentioned in more detail later under the week, especially under "Archiving principles".

Wednesday. Metadata and interoperability

The Wednesday sessions were entirely dedicated to metadata and interoperability. Mr. Uneson introduced the problem of discovering and managing digital resources in large collections and pointed at metadata, structured machine-readable answers to the "who/what/where/when/how" of a digital object, as a partial solution. He mentioned the concepts of descriptive/administrative/structural and external/embedded metadata and pointed out that a metadata specification is independent of the syntax it is expressed in.

After having given examples of possible usages of metadata, such as searching, browsing, automatically accessing appropriate tools, ingesting new resources, etc, Mr. Uneson touched common problems: a) full consistency is difficult to achieve for metadata generated by humans; b) pre-defined metadata sets may be too inflexible for a given purpose; c) all metadata definitions need to strike a balance between lower precision/lower functionality on one side and higher workload/lower interoperability on the other, a balance sometimes difficult to find. He suggested that good tools and metadata sets which allow for user-defined extensions may be partial solutions to all of these issues.

As an example of a small and very general metadata set, Mr. Uneson then introduced Dublin Core[22] (DC), primarily a librarian's tool and meant for non-specialist creators and users[23]. In DC, all elements are optional and repeatable. Mr. Uneson briefly introduced DC qualifiers to express encoding or refinement, the "One-to-one Principle", and the "Dumb-down-principle". He concluded with a practical demonstration of one web-based[24] and one stand-alone[25] tool for producing Dublin Core in desired syntax. Mr. Uneson then gave examples of much larger and therefore more work-intensive metadata sets. First, he briefly mentioned the Text Encoding Initiative[26] (TEI), a combined metadata and annotation scheme for literary and linguistic texts with full specification available in XML. Thereafter, he went on to describe the International Standards for Language Engineering (ISLE) Metadata Initiative (IMDI) in some detail.

IMDI is a domain-specific metadata set for language resources, intended to be sufficiently rich for professional purposes and meant to work in a distributed domain where the physical location of

data is of no concern to the user. Currently, IMDI is reasonably mature and stable for multimedia resources; written resources are also supported, but the accompanying vocabularies are less stable. User-defined keys are allowed as a complement to the predefined descriptors. In the IMDI approach, metadata is always external to the data it describes, pointing to it via http links; while the metadata itself always is open, this physical separation of metadata and (distributed) data also allow for convenient restriction of access. An IMDI file may group any number of related resources, perhaps a primary audio/video recording, a transcription, a translation, some field notes or photos, etc; or (for written resources) a primary text, a few translations, some comments, illustrations, statistical analyses, etc. Apart from searching, IMDI files may also be grouped in arbitrary hierarchies to permit browsing, and any file found may be started with the tool of the user's preference, if the system is appropriately configured.

Mr. Uneson then concluded by demonstrating the main IMDI tools, the IMDI editor and the BC (Browsable Corpus) browser. The IMDI editor provides a user-friendly GUI for generating metadata description files, including guidelines for using the metadata vocabulary. It restricts the input of values where needed and allows the user to save and reuse blocks of data. The editor permits automatic downloading of configurable controlled vocabularies and groups of such vocabularies, known as profiles. This arrangement permits a smaller community to define descriptors for their particular needs or projects, as have indeed been done by researchers in Sign Language. The BC browser, which he thereafter showed, may be used to browse and search the metadescriptions thus generated, and the identified resources may be immediately downloaded or otherwise exploited.

Mr. Wittenburg began the afternoon sessions by talking about the goal of interoperability as the possibility of joint operations on data from different sources, permitting transparent use of tools and operations on different data types (currently textual data only). He explained that making structure explicit alone (as we might with for instance XML) is not enough to achieve interoperability -- identical semantic relations may be expressed in different structures, and identical structures do not necessarily carry identical semantics. What we need, he explained, are formal frameworks to express ontologies, data categories and relations between them, which makes interoperability a question of ontology matching.

Such formal frameworks are indeed under development. Mr. Wittenburg mentioned the view of the work of ISO 11179[27], where domain-meaningful data categories (data element concepts) and their corresponding values are given language- and implementation-independent abstract labels such as /Gender/ and {/masculine/, /feminine/, /neuter/}, respectively. In more detail, he then touched the W3C[28] view with its stack of languages, in which the syntactically oriented standards XML[29] and XML-Schema[30] are complemented by the W3C recommended standards

- Resource Description Framework (RDF)[31]: a language for making statements about web resources, statements of the type Object--Attribute-->Value triples, where objects are web resources. All identifiers may be URLs; allowing for www-unique naming; values can also be objects, as can entire RDF statements, permitting nested statements (e.g., "According to X, Y [is] authorOf Z")..

- RDF-Schema[32]: a language permitting definitions of a vocabulary for RDF and organizing this vocabulary in a typed hierarchy.

- Web Ontology Language (OWL[33]): a complete knowledge representation language, permitting class/individual equality, inequality, several logical operations, etc.

Mr. Wittenburg then went on to describe the case of interoperability in ECHO, where the interoperability of many different metadata sets[34] had to be solved somewhat less elegantly, due to lack of time, tools, and reliable ontologies. In ECHO, the ontology mapping[35] partially had to be performed manually, and the mapping had to include fuzzy relations, especially "mapsTo". Such relations cannot be used in binary logic and are thus of limited usefulness for non-fuzzy inference engines. He concluded that a working RDF/OWL domain still is rather far away. Furthermore, a further prerequisite for such a domain is some mechanism for defining and registering relations in a repository, still an unsolved issue.

Going on with the ECHO example, Mr. Wittenburg then demonstrated the Digital Open Resource Area (DORA) interface[36]: its facilities for browsing (where available); geographic selection (where available); complex search on a domain-specific metadata set (for specialists, taking advantage of the full expressivity of the current metadata set); and Google-like full-text search (for non-specialists, permitting unstructured search on any element, less powerful but not requiring any particular knowledge from the user).

Thursday. Archiving principles. Workflow. File formats

Mr. Wittenburg began the Thursday sessions by a talk on general principles for digital archiving. He returned to the important theme "representation versus presentation" from the introduction. Researchers often have a short-time perspective, tend to focus on interpretation and analysis of collected data and a certain design, and use a certain set of short-lived tools to do this. While this view is certainly needed to achieve any results, it is not the most efficient one for long-term archiving with focus on data reusability and preservation -- instead, it is more fruitful to adopt the archivist's view, honouring open, documented formats and standards and striving for independence of the currently available tools. From such a representation, the generation of particular presentations can be regarded as something like an export option. For a concrete example, he pointed at open XML representations as superior for long-term preservation to presentation formats such as HTML or proprietary tool formats such as those produced by many commercial relational databases (MS Access, Filemaker).

After having reminded the audience of the important life-time aspects of any archive and the necessity to find working models for physical data storage, both touched somewhat in the introductory talks, Mr. Wittenburg then went on to the infrastructure of an archive. First, he turned to the question of data containers, and compared three common approaches: to have all resources directly addressable in a file system; or to have them managed by a database manager (DBMS) shell; or by a Content Management System (CMS), originally intended to help managing web sites. The file system is the simplest by far, with low learning and buying costs associated; the user may access every bit and there is no dependence on external companies. On the other hand, hierarchies which do not follow file system hierarches may be less easily managed. Of course, shells may be constructed around a file system anyway, creating hybrid solutions; the point is that the data is directly adressable as well. DBMS solutions may provide appealing user interfaces, but there is a risk of inflexibility and dependence on external entities. Furthermore, for relational DBMS (rDBMS), the currently by far most popular design, the Entity-Relation data model is applied on any data, and this may or may not be appropriate. CMS on the other hand is more flexible but the learning costs may be higher. It may be combined with the file system approach. There are commercial as well open source solutions for both RDBMS (e.g., MySQL[37], PostgreSQL[38]) and CMS (e.g., ZOPE[39]). There are tradeoffs involved with any solution; typical management tasks such as copying, moving, modifying, versioning, consistency checking, and setting up services will be rather differently performed if done, say, script-wise by the institute's own archive or system manager, or via a predefined user-interface that an external company may have provided against payment. The latter solution may provide fast results, but when employing external services with the long-term perspective necessary for most archiving work, it is wise observe a general concern and be prepared for the possibility that the service becomes much more expensive, or disappears.

Going on with archiving principles, Mr. Wittenburg made a difference between universal archival issues on one side, such as principles of physical survival, choice of structuring language (XML), and basic encoding formats (such as UNICODE for text; TIFF/PNG for graphics; MPG for video), and discipline specific on the other, such as presentation design, tools, interpretations, etc.

Turning to workflow, he then described of the DoBeS project, with a per-team contract about which media formats, file formats, file naming systems, ways of interaction, integration, timetables to use; how to identify file groupings, intermediate and versions, etc. The current media formats accepted in the DoBeS project are DV, VCD, Hi-8, VHS, S-VHS, NTSC/PAL for video, and Uher 4400, DAT, MD, CR for audio. While legacy hardware and practical considerations may be deciding, some formats (DV, Uher 4400) are preferable to others, for lack of quality (e.g., VHS) or uncertain hardware supply (e.g., DAT).

The speakers then went on to formats used for archiving. Mr. Uneson talked about the general archiving principle of having one, use-neutral master file in a format that does not restrict future options more than absolutely necessary. This "Rich Digital Master" may then be parent to many children, each suited for a certain purpose, such as printing, www publishing, multimedia productions, different kinds of analysis, etc. Thereafter, he went on to archiving text, and gave examples of different levels of digitization. Page images may be crucial for certain material, such as mediaeval manuscripts, but of less interest in others, such as recent publications. Page images may be accompanied by textual representations, produced by OCR with or without proofreading, or where OCR is impossible, by keying. The possibilities require different skills and imply different levels of workload, from almost automatic to very high. The balance between workload and usefulness must be set separately for each project. However, workflow-wise, enhancements could be done later: a raw OCR scan linked to page images can later be proofread.

Digital texts may be made more useful by adding markup, different means of making explicit a certain interpretation of a text. However, much of markup work must be made manually, again making a balance between workload and usefulness. Markup is omnipresent in web standards such as HTML. The versatile XML metalanguage, used to defined specific markup languages, has been used for many projects within Cultural Heritage, such as the previously mentioned TEI.

He then turned to image formats. There are many different ones, many of which are invented by software companies. Some of the most popular ones have evolved to de-facto standards, in particular TIFF[40]. There are no perfect solutions in such cases; while challenging the idea of using open standards for archive material, de-facto standards may be difficult to avoid. TIFF is widely supported and used, and it is probably a better choice than an open but little supported standard. An open alternative deserving consideration, however, is the PNG[41] image file format (ISO 15948), spotting lossless compression. Lossy compressed images (such as JPEG[42]) generally should be avoided as archiving formats, as should proprietary image formats for editors, however popular (PSP, PSD, CPT); proprietary image formats for operative systems (BMP, PICT); and printing-oriented page description file formats (PS, EPS).

Mr. Wittenburg concentrated on media file compression. He described the general idea underlying (lossy) compression, that not all information present in the signal is equally important, and that much of it will not be perceived by human beings anyway. For some purposes, it may thus be deleted. Uncompressed video typically uses about 250 Mbps (about 100 GB/h); with current storage media and bandwidths, compression is of course unavoidable. Also for audio (0.8 - 2.3 Mbps, 350 - 1000 MB/h), compression is tempting. Mr. Wittenburg showed the results of an experiment testing a few speech-related algorithms on uncompressed and speech compressed with different techniques, with very small differences in the outcome. However, as he pointed out, once we have decided on compression, the discarded information is irrevocably lost. As we do not know what future analyses and techniques there may be, and whether they will be dependent on the discarded information, we might want to try to capture as much as possible when we can. The fast development of storage capacity has meant that audio can now generally be stored uncompressed; it might be that we will even use uncompressed video for archiving in some years from now.

The speakers then summarized current recommended file format practices. For text, the UNICODE encoding is desirable, as is XML for any markup, if possible reusing some existing standard. For images, TIFF or PNG are recommended. Typical colour depth/resolution choices, which however should be considered particularly for a given project, are 24 bit colour/300dpi or 8 bit greyscale/600dpi.

The current best master file practice for audio encoding is non-compressed 16 bit/48kHz (0.8 Mbps, 350 MB/h) or 24 bit/96kHz (2.3 Mbps, 1GB/h) linear PCM; formats often used are WAV, AIFF, or NIST. For delivery, the compressed MP3 is currently common. As for video, the MPEG2 encoding is often used (6 Mbps, 3GB/h) for archiving purposes; SMIL, AVI, and MPEG are typical formats. For delivery, formats with more compressed types such as MPEG1 (1.5 Mbps, 700MB/h) and MPEG4 (0.5 Mbps, 200 MB/h) are often used.

Friday. Content. Archive management. Access management

The first session on Friday was dedicated to content analysis and annotation. Once a new resource is created, or perhaps created earlier and discovered by metadata, it might be appropriate to annotate it for future use, and/or to analyse it as it is, for immediate exploitation. A few annotation and/or analysis tools for different data types were demonstrated by Mr. Wittenburg and Mr. Uneson:

- Annotea[43], a W3C project within its Semantic Web activities, permits annotations (comments, notes, explanations, etc) to be attached to any web document or part of it in a distributed fashion. The annotations are saved on one or more dedicated annotation servers; there is no need to without actually touch the annotated document itself. Annotea is open and uses important W3C standards such as RDF and XPointer. Annotea is implemented in the W3C showcase editor/browser Amaya[44].

- Transcriber[45], a free and open (GPL) tool for assisting the manual annotation of speech signals, providing a graphical user interface for segmenting, transcribing, labeling long duration speech recordings, especially useful for the annotation of broadcast news recordings and similar.

- Praat[46], a free and powerful tool for speech analysis, synthesis, and manipulation, written by Paul Boersma and David Weenink at the Department of Phonetics of the University of Amsterdam, available for most platforms and as source code. While less useful for annotation, most analyses a phonetician may wish to have are implemented.

- ELAN[47], Eudico Linguistic Annotator, a free and open (GPL), Java-based tool used to annotate multimedia files of practically any length, permitting convenient annotations of different types (such as symbolic or time-aligned) with Unicode support in an arbitrary number of tiers, and offering many different visualizations of annotations.

Mr. Wittenburg then went on to discuss archive management issues. He gave an overview of the situation at MPI-PL, where the two main positions are one Archive Manager, controlling workflow and leading digitization and integration, and one Technical Corpus Manager, responsible for management and consistency (but with no direct influence on content) He emphasized the importance of treating all persons involved, from data creators and digitizers to corpus managers and system administors, as equal players in the team.

The MPI-PL approach, where all tasks are performed locally, is a fairly traditional and well-tested setup in a European perspective. An interesting alternative, made possible rather recently with increasing bandwidths, is to outsource tasks that are not dependent on actual archive content. Mr. Wittenburg mentioned the case of the Australian Paradisec archive, where the responsibility of technical and system tasks such as backup, security, and consistency checks have been entirely handed over to an external National Computer Centre for archiving and IT, over a very fast connection. In this scenario, project members may concentrate locally on what they are best at, such as PR and archive management (digitization, metadata creation, workflow, etc). The need for expensive archiving hardware and staff thus decreases.

Going on, Mr. Wittenburg turned to access management. He concentrated on the large-scale scenario, where all users are seen as one community, consisting of many sub-groups, and where each sub-group has a unique name and is associated with certain access rights. Rights should probably be time-limited by default; at any rate, a user of a large archive should sign some well-documented, formal agreement, such as a Code of Conduct (CoC). He gave the ethical and legal considerations in the DoBeS project as an example[48] and showed the CoC used there[49]. Mr. Wittenburg then enumerated some important questions in this scenario:

a) what rights should a user have?
b) who can/should give these rights?
c) who can/should give the rights to thus define rights?
d) how could rights efficiently be defined at archive, corpus, resource level?

e) how should the right-giving system be managed efficiently?

For the first question, defining reading rights is unproblematic (however, as Mr. Wittenburg reminded the audience, there is no easy way to completely rule out abuse through copying, although watermark techniques for media and image files may decrease this risk). Writing access is more controversial, since it may destroy archive consistency.

For the remaining questions, tractability is an important aspect. Any fully centralized solution is not scalable -- when the numbers of users and data items with different access restrictions become high enough, the workload soon becomes too large. In a distributed system access rights management must be delegated. Mr. Wittenburg described the current solution at the Language Resource Archive of MPI-PL, where each sub-corpus generally has a responsible researcher who is authorized to modify access rights for that corpus (but no others) via a web interface. All access is defined in terms of user groups, and a user may be a member of any number of groups. The web interface now permits an authorized person to add individual users; to add/modify/delete groups; to set what kind of access restriction should apply for a given group at a given node in the subcorpus; and finally, to add other authorized users, having the same permissions as the already authorized person.

Mr. Wittenburg then briefly sketched a project called DAM-LR, still in its early phase, involving MPI-PL, Lund University, SOAS[50] (School of Oriental and African Studies, University of London, UK), and INL[51] (Institute for Dutch Lexicology, Leiden, the Netherlands). DAM-LR wants to integrate techniques for unique resource identification and resolving (based on the American CNRI [Corporation for National Research Initiatives] Handle[52] system), distributed user/group and access management system (based on the American university consortium Internet2 Shibboleth[53] system) with the existing distributed IMDI metadata domain. This scenario would offer many advantages to the current; first and foremost, a user could get at any resource to which he has been granted access rights, irrespective of its physical location, with one and only one user ID. Mr. Wittenburg concluded by giving a fairly technical description of the current storage setup and the hardware employed at MPI-PL.

Digitizing Intangible Cultural Heritage--Analysis and recommendations

Marcus Uneson
Department of Linguistics, Lund University
Lund
Sweden

Peter Wittenburg
Max Planck Institute for Psycholinguistics
Nijmegen
The Netherlands

This section summarizes the authors' impressions of the state of ongoing cultural heritage digitization activities in Lithuania and tries to point at some areas where we think it would be advantageous to consider alternative approaches. The section is part of the authors' report from the course, but phrased to be self-contained and thus possible to circulate separately, should this be convenient. For fuller details, see the rest of the report, as well as slides, links, etc, on the course web site

http://www.ling.lu.se/projects/echo/events/vilnius/

Introduction

As part of Phase III of the project "Establishment of a National Inventory and Electronic Database of Lithuanian Intangible Cultural Heritage", founded by UNESCO, the authors offered a course in digital archiving called "Digitizing Intangible Cultural Heritage" in Vilnius, Lithuania, March 15 to 20, 2004.

The contacts were established via Lund University as being a member of the current EU-funded project ECHO (European Cultural Heritage Online, 18 months, Dec 2002 -- May 2004), an international endeavour with some fifteen European partners and led by three Max Planck Institutes. The main goal of ECHO is the establishment of a European infrastructure fostering the transfer of cultural heritage to the internet, permitting free access to fully interoperable, standards-compliant corpora of primary cultural heritage documents, as well as tools to exploit these documents. For more on ECHO, see

http://echo.mpiwg-berlin.mpg.de/home
(ECHO Max Planck Institute for the History of Science, main site)

http://www.ling.lu.se/projects/echo/contributors/
(ECHO Lund, Language resource contributors)

http://www.mpi.nl/echo/
(ECHO Max Planck Institute for Psycholinguistics, Technical infrastructure)

Representatives of several institutions for Cultural Heritage of Lithuania participated in the course. To the authors, this meant an excellent opportunity to get a glimpse of the current state of the construction of digital cultural heritage archives in Lithuania.

The authors were impressed by the amount of activities to take care about Lithuanian history, to revitalize old traditions, and to document them. In this respect, Lithuania seems to be more active than many Western countries and has little to learn from them. Rather, it was we who learned --

we tried to bring home a bit of the enthusiasm we met. Just to mention one of the many interesting projects which were introduced to us, the scholarly program "Expressions of Lithuanian Mental Culture: The Ethnological, Linguistic, and Historical Database" (http://www.aruodai.lt; at the time of writing, very little information in English) appears as a very interesting cross-disciplinary joint enterprise, launched by the Institute of Lithuanian Literature and Folklore, the Institute of the Lithuanian Language, the Lithuanian Institute of History, and the Institute of Mathematics and Informatics. Its aim is a comprehensive electronic collection of sources in language, folklore, ethnology, archeology, and history.

We were likewise pleasantly surprised by the excellent presentation work. Representatives of the different participating institutions showed us many well-designed CDs, publications, editions, web-sites, etc, and much of this work exploits the current state-of-art possibilities.

Analysis

In the following, we comment a few recurring issues of the week, where we have particular recommendations or points to make.

presentation versus representation	The high priority given to presentation aspects at this moment is understandable, since it delivers products for the public, researchers, schools and others. Presentation work, however, is inherently focused on short-term aspects and therefore associated with choices that can lead to severe problems when thinking about long-term digital preservation of the material, such as aspects of encoding, formatting, structuring and container types. Here we felt that the course was an excellent opportunity to work out the differences between presentation and representation, and to create an awareness about the needs for long-term preservation.
independence and individualism of institutes	The participants mentioned the independence and individualism of the various institutions. While a certain amount of competition may be instrumental to stimulate creative atmospheres, leading perhaps to products such as the ones mentioned, this is not the most suitable form for achieving long-term goals. For instance, each institute faces very similar or identical questions when it comes to physical data storage and backup.
metadata, general	Metadata is a general concern for digital archives, well-known to many institutions and museums world-wide. While it does not solve all problems, it supports management and discovery in continuously growing digital collections. In the course, metadata issues raised much more interest than we had expected. We can only hope to have created an awareness of its advantages, as working solutions need some time to mature.
metadata, interoperability	Interoperability of metadata also raised intensive discussions and is indeed seen as increasingly important world-wide. There are no perfect solutions yet, but frameworks for metadata interoperability are under development. The core pillars of interoperability in the future will be data category registries (for instance, those planned by ISO TC37/SC4 for the domain of language resources), which define concepts in an open and machine readable way, and frameworks such as RDF and OWL.
media and strategies for long term preservation	We found at times a lack of awareness that long-term preservation cannot be done with the help of individualized data and storage media. For example, the creation of CDROMs may be excellent to serve the current needs and to produce presentations; however, these storage media will degrade within a very short time. The authors believe that only continuous migration and massive copying will increase the chances of data survival. Here we see a lack of infrastructure in Lithuania – at least this is our impression from the many talks we had with the participants.

Recommendations

1. We recommend that projects and institutes should look carefully at the infrastructures that have already been worked out by other groups. Slight modifications may help to carry out the necessary work. First and foremost, the institutions have to describe the metadata elements that would fit their actual needs. Thereafter, they may look at existing standards to see whether these may be adapted with reasonable effort. The advantage of such a procedure is that existing infrastructures and tools may be re-used. Since metadata creation and maintenance can be very labour-intensive, this is an important point. There are useful frameworks for metadata already now; however, only the emerging data category registers and RDF techniques will allow all of the desired flexibility.
   The techniques mentioned will provide important frameworks for interoperability in the future. In the meantime, we recommend the adherence to standards such as XML and OAI-PMH, in order to facilitate later data integration into wider efforts.

2. We see a need to establish a "Centre for Modern IT concepts for the Humanities". The existing Institute for Mathematics and Informatics may provide a good starting point. Such a centre does not have to provide services for the presentation aspects -- this knowledge seems already to be widely spread. Instead, it should focus on methodologies to build up digital collections based on open standards for long-term preservation, to select appropriate containers, to bridge the gap between presentation and representation at all levels, to tackle the discussed metadata and interoperability problems, and to meet the requirements of the emerging Semantic Web. It has to provide services, i.e., it has to do joint projects with cultural heritage institutions, give training courses, and offer support.
   These task descriptions indicate that such a centre should have staff that carries out their own methodological research and that participates in international circuits. However, the centre must be closely in touch with the work in the cultural heritage institutions.

3. Further, we recommend the establishment of a "National Computing Centre for Cultural Heritage", where all cultural heritage material emerging from the work done in the different cultural heritage institutions is stored. Such a centre could and should employ techniques of continuous migration and mass storage, generally no viable solutions for individual cultural heritage institution. In this way, many of the similar or identical questions regarding physical data storage and backup that each of the institutions is facing now are solved centrally, more reliably and more efficiently.
   There should be a formal agreement specifying that copies of all digital cultural heritage material be sent to this centre. The centre can be housed at a more general National Computing Centre, if such an institution already exists. Outside the task of providing long-term storage, this centre would link up with European Data Grid initiatives. It would also need to collaborate very closely with the Centre for Modern IT concepts suggested above.

Conclusion

We would like to add that we were charmed by the intensity of the discussions we had, and by the enthusiasm of the participants. We see this enthusiastic attitude as an excellent basis for further plans in the directions mentioned. If Lithuania in a short time frame could establish infrastructures as those described, we imagine that the country could serve as an example for other nations in comparable situations.

Links

[1] http://www.unesco.org
[2] http://echo.mpiwg-berlin.mpg.de/home (ECHO MPI-WG)
http://www.ling.lu.se/projects/echo/contributors/ (ECHO Lund)
http://www.mpi.nl/echo/ (ECHO MPI-PL)
[3] http://www.lfcc.lt/
[4] http://www.unesco.org/culture/heritage/intangible/
[5] http://www.lfcc.lt
[6] http://www.unesco.mii.lt
[7] http://www.llti.lt
[8] http://www.lma.lt
[9] http://www.istorija.lt
[10] http://www.mab.lt
[11] http://www.lki.lt/English/index.html; example at http://www.mch.mii.lt/dba/index.htm
[12] A few articles in Lithuanian:
http://www.ebiz.lt/article.php3/15/6095/6
http://www.aruodai.lt/c_kronika.htm
[13] http://www.ling.lu.se/projects/echo/contributors/events/vilnius/index.html
[14] http://www.ling.lu.se/projects/echo/contributors/events/vilnius/index.html
[15] http://www.mpi.nl
[16] http://www.ling.lu.se
[17] http://nausikaa2.mpiwg-berlin.mpg.de:86/cgi-bin/toc/toc.x.cgi?dir=monan_mecha_035_la_1599&step=thumb
[18] http://nausikaa2.mpiwg-berlin.mpg.de:86/cgi-bin/toc/toc.x.cgi?dir=buch_atlas_fr_01_1836&step=thumb
[19] http://www.ling.lu.se/persons/Marcusu/echo/state_of_art/
[20] http://www.sol.lu.se/
[21] http://www.nyu.edu/its/humanities/ninchguide/index.html
[22] http://dublincore.org/
[23] http://dublincore.org/documents/usageguide/
[24] http://www.ukoln.ac.uk/metadata/dcdot/
[25] http://metabrowser.spirit.net.au/. Commercial, free for DC.
[26] http://www.tei-c.org/
[27] http://www.iso.org; search for ISO 11179:1-6 (1995-2003)
[28] http://www.w3.org/
[29] http://www.w3.org/XML/
[30] http://www.w3.org/XML/Schema
[31] http://www.w3.org/RDF/
[32] http://www.w3.org/TR/1999/PR-rdf-schema-19990303/
[33] http://www.w3.org/2004/OWL/
[34] http://www.mpi.nl/echo/tec-rep/wp2-tr16-2003v3.pdf
[35] http://www.mpi.nl/echo/tec-rep/wp2-tr17-2004v2.pdf
[36] http://www.mpi.nl/echo/tec-rep/wp2-tr18-2004v2.pdf
[37] http://www.mysql.com
[38] http://www.postgresql.org
[39] http://www.zope.org
[40] http://home.earthlink.net/%7Eritter/tiff (unofficial TIFF home page)
[41] http://www.libpng.org/pub/png/
[42] http://www.jpeg.org/public/jpeglinks.html
[43] http://www.w3.org/2001/Annotea/
[44] http://www.w3.org/Amaya/
[45] http://www.etca.fr/CTA/gip/Projets/Transcriber/
[46] http://www.fon.hum.uva.nl/praat/
[47] http://www.mpi.nl/tools/elan.html
[48] http://www.mpi.nl/DOBES/INFOpages/applicants/legal-ethics-issues.html
[49] http://www.mpi.nl/DOBES/INFOpages/legal_ethic/codeOFconductNew.html
[50] http://www.soas.ac.uk/
[51] http://www.inl.nl/

[52] http://www.handle.net
[53] http://shibboleth.internet2.edu/

Appendix A. List of participants[1]

Institute of Mathematics and Informatics

Prof. Dr. Habil. Laimutis Telksnys
Head of the UNESCO Chair in Informatics for the Humanities
E-Mail: telksnys@ktl.mii.lt

Dr. Nerute Kligiene
Project Leader of the UNESCO Chair in Informatics for the Humanities
E-Mail: nerute@ktl.mii.lt

Evaldas Ozeraitis
Consultant, programmer of the UNESCO Chair in Informatics for the Humanities
E-Mail: evaldas@ktl.mii.lt

Institute of Lithuanian Literature and Folklore

Department of Folk Songs

Dr. Brone Stundziene
Assistant Director,
Head of Department of Folk Songs
E-Mail: brone@llti.lt

Department of Folklore Archives

Dr Ruta Žarskiene
Head of Department
E-Mail: ruta@llti.lt

Dr. Auste Naikene
Ethnomusicologist
E-Mail: auste@llti.lt

Department of Folk Narrative

Dr. Daiva Vaitkeviciene
Scientist
E-Mail: vaidaiva@llti.lt

[1] This list is only an unofficial update of the preliminary list of participants and so may contain occasional errors. Reported errors will be corrected on the list on the web site of the course, http://www.ling.lu.se/projects/echo/contributors/events/vilnius/

Lithuanian Music Academy

Department of Ethnomusicology

Assoc. Prof., Dr. Daiva Vyciniene
Head of department
E-mail: daiva.r@is.lt

Institute of Musicology/Section of Ethnomusicology

Dr. Dalia Urbanaviciene
Head of section
E-mail: daliau@delfi.lt

Rytis Ambrazevicius
Scientist
E-mail: pvs@takas.lt

Dr. Gaila Kirdiene
Scientist
E-mail: daliau@delfi.lt

Audio, Video and Internet centre
Antanas Auskalnis
Director
E-mail: antanas.auskalnis@lma.lt

Lithuanian Institute of History

Department of Ethnology

Dr. Zilvytis Bernardas Saknys
Head of Department
E-mail: zilvytis@takas.lt

Department of Archeology

Dr. Vykintas Vaitkevicius
Scientist
E-mail: vikivait@takas.lt

Lithuanian Folk Culture Centre

Department of Ethnic Culture

Vida Satkauskiene
Head of Department
E-mail:lfcc@lfcc.lt

Sigitas Jonkus
Consultant in Database Projecting
E-mail: sigitas.j@takas.lt

Section of Folklore

Jurate Semetaite
specialist
E-mail: lfcc@lfcc.lt

Arunas Lunys
specialist
E-mail: a.lunys@lfcc.lt

Audrone Vakariniene
specialist
E-mail: folkloras@lfcc.lt

Loreta Sungailiene
specialist
E-mail: l.mukaite@lfcc.lt

Section of Folk Art

Terese Jurkuviene
Head of section
E-mail: tautodaile@lfcc.lt

Dr. Elena Pociulpaite
specialist
E-mail: tautodaile@lfcc.lt

Section of Information of Ethnic Culture

Inga Krisciuniene
Head of section
E-mail: ekinfo@lfcc.lt

Dalia Usinaviciene
specialist
E-mail: lfcc@lfcc.lt

Archive of Folk Culture
Rita Balkute
Archyvist
E-mail: archyvas@lfcc.lt

Video, audio , photo laboratory

Darius Linge
Specialist of technical equipment
E-mail: lfcc@lfcc.lt

Egidijus Baniunas
Video operator
E-mail: videolab@lfcc.lt

Ramunas Virkutis
Photograph
E-mail: fotolab@lfcc.lt