Mockus, Audris. Roy T. Fielding, James D. Herbsleb. "Two Case Studies of Open Source Software Development: Apache and Mozilla." *ACM Transactions on Software Engineering and Methodology* 11.3 (2002): 309-46.

Skinner, David C. *Introduction to Decision Analysis*. 2nd ed. Sugar Land, TX: Probabilistic, 1999.

Aspects of Sustainability in Digital Humanities

Georg Rehm
georg.rehm@uni-tuebingen.de
Tübingen University, Germany

Andreas Witt
andreas.witt@uni-tuebingen.de
Tübingen University, Germany

The three papers of the proposed session, "Aspects of Sustainability in Digital Humanities", examine the increasingly important topic of sustainability from the point of view of three different fields of research: library and information science, cultural heritage management, and linguistics.

Practically all disciplines in science and the humanities are nowadays confronted with the task of providing data collections that have a very high degree of sustainability. This task is not only concerned with the long-term archiving of digital resources and data collections, but also with aspects such as, for example, interoperability of resources and applications, data access, legal issues, field-specific theoretical approaches, and even political interests.

The proposed session has two primary goals. Each of the three papers will present the most crucial problems that are relevant for the task of providing sustainability within the given field or discipline. In addition, each paper will talk about the types of digital resources and data collections that are in use within the respective field (for example, annotated corpora and *syntactic treebanks* in the field of linguistics). The main focus, however, lies in working on the distinction between field-specific and universal aspects of sustainability so that the three fields that will be examined – library and information science, cultural heritage management, linguistics – can be considered case studies in order to come up with a more universal and all-encompassing angle on sustainability. Especially for introductory texts and field – *independent* best-practice guidelines on sustainability it is extremely important to have a solid distinction between universal and field-specific aspects. The same holds true for the integration of sustainability-related informational units into field-independent markup languages that have a very broad scope of potential applications, such as the TEI guidelines published by the Text Encoding Initiative.

Following are short descriptions of the three papers:

The paper "Sustainability in Cultural Heritage Management" by Øyvind Eide, Christian-Emil Ore, and Jon Holmen discusses technical and organisational aspects of sustainability with regard to cultural heritage information curated by institutions such as, for example, museums. Achieving organisational sustainability is a task that not only applies to the staff of a museum but also to education and research institutions, as well as to national and international bodies responsible for our common heritage.

Vital to the sustainability of collections is information about the collections themselves, as well as individual items in those collections. "Sustaining Collection Value: Managing Collection/ Item Metadata Relationships", by Allen H. Renear, Richard Urban, Karen Wickett, Carole L. Palmer, and David Dubin, examines the difficult problem of managing collection level metadata in order to ensure that the context of the items in a collection is accessible for research and scholarship. They report on ongoing research and also have preliminary suggestions for practitioners.

The final paper "Sustainability of Annotated Resources in Linguistics", by Georg Rehm, Andreas Witt, Erhard Hinrichs, and Marga Reis, provides an overview of important aspects of sustainability with regard to linguistic resources. The authors demonstrate which of these several aspects can be considered specific for the field of linguistics and which are more general.

Paper 1: Sustainability in Cultural Heritage Management

Øyvind Eide, Christian-Emil Ore, Jon Holmen
University of Oslo, Norway

Introduction

During the last decades, a large amount of information in cultural heritage institutions have been digitised, creating the basis for many different usage scenarios. We have been working in this area for the last 15 years, through projects such as the Museum Project in Norway (Holmen et al., 2004). We have developed routines, standardised methods and software for digitisation, collection management, research and education. In this paper, we will discuss long term sustainability of digital cultural heritage information. We will discuss the creation of sustainable digital collections, as well as some problems we have experienced in this process.

We have divided the description of sustainability in three parts. First we will describe briefly the technical part of sustainability work (section 2). After all, this is a well known research area on its own, and solutions to many of the problems at hand are known, although they may be hard to implement. We will then use the main part of the paper to discuss what we call organisational sustainability (section 3), which may be even more important than the technical part in the future — in our opinion, it may also be more difficult to solve. Finally, we briefly address the scholarly part of sustainability (section 4).

Technical Sustainability

Technical sustainability is divided in two parts: preservation of the digital bit patterns and the ability to interpret the bit pattern according to the original intention. This is an area where important work is being done by international bodies such as UNESCO (2003), as well as national organisations such as the Library of Congress in the USA (2007), and the Digital Preservation Coalition in the UK (DPC,2001).

It is evident that the use of open, transparent formats make sit easier to use preserved digital content. In this respect XML encoding is better compared to proprietary word processor formats, and uncompressed TIFF is more transparent than company-developed compressed image formats. In a museum context, though, there is often a need to store advanced reproductions of objects and sites, and there have been problems to find open formats able to represent, in full, content exported from proprietary software packages. An example of this is CAD systems, where the open format SVG does not have the same expressive power as the proprietary DXF format (Westcott, 2005, p.6). It is generally a problem for applications using new formats, especially when they are heavily dependent upon presentation.

Organisational Sustainability

Although there is no sustainability without the technical part, described above, taken care of, the technical part alone is not enough. The organisation of the institution responsible for the information also has to be taken into consideration.

In an information system for memory institutions it is important to store the history of the information. Digital versions of analogue sources should be stored as accurate replica, with new information linked to this set of historical data so that one always has access to up-to-date versions of the information, as well as to historical stages in the development of the information (Holmenetal.,2004, p.223).

To actually store the files, the most important necessity is large, stable organisations taking responsibility. If the responsible institution is closed without a correct transfer of custody for the digital material, it can be lost easily. An example of this is the Newham Archive (Dunning, 2001) incident. When the Newham Museum Archaeological Service was closed down, only a quick and responsible reaction of the sacked staff saved the result of ten years of work in the form of a data dump on floppies.

Even when the data files are kept, lack of necessary metadata may render them hard to interpret. In the Newham case the data were physically saved but a lot of work was needed to read the old data formats, and some of the information was not recoverable. Similar situations may even occur in large, stable organisations. The Bryggen Museum in Bergen, Norway, is a part of the University Museum in Bergen and documents the large excavation of the medieval town at the harbour in Bergen which took place from the 1950s to the 1970s. The museum stored the information in a large database.

Eventually the system became obsolete and the database files were stored at the University. But there were no explicit routines for the packing and future unpacking of such digital information. Later, when the files were imported into a new system, parts of the original information were not recovered. Fortunately all the excavation documentation was originally done on paper so in principle no information was lost.

Such incidents are not uncommon in the museum world. A general problem, present in both examples above, is the lack of metadata. The scope of each database table and column is well known when a system is developed, but if it is not documented, such meta-information is lost.

In all sectors there is a movement away from paper to born digital information. When born digital data based on archaeological excavations is messed up or lost – and we are afraid this will happen – then parts of our cultural heritage are lost forever. An archaeological excavation destroys its own sources and an excavation cannot be repeated. For many current excavation projects a loss of data like the Bryggen Museum incident would have been a real catastrophe.

The Newham example demonstrates weak planning for negative external effects on information sustainability, whereas the Bergen example shows how a lack of proper organisational responsibility for digital information may result in severe information loss. It is our impression that in many memory institutions there is too little competence on how to introduce information technologies in an organisation to secure both interchange of information between different parts of the organisation and long-term sustainability of the digital information. A general lack of strategies for long term preservation is documented in a recent Norwegian report (Gausdal, 2006, p.23).

When plans are made in order to introduce new technology and information systems into an organisation one has to adapt the system to the organisation or the organisation to the system. This is often neglected and the information systems are not integrated in the everyday work of the staff. Thus, the best way to success is to do this in collaboration and understanding with the employees. This was pointed out by Professor Kristen Nygaard. In a paper published in 1992 describing the uptake of Simula I from 1965 onwards, Nygaard states: "It was evident that the Simula-based analyses were going to have a strong influence on the working conditions of the employees: job content, work intensity and rhythm, social cooperation patterns were typical examples" (Nygaard, 1992, p. 53). Nygaard focused on the situation in the ship building industry, which may be somewhat distant from the memory institutions. Mutate mutandis, the human mechanisms are the same. There is always a risk of persons in the organisation sabotaging or neglecting new systems.

Scholarly Sustainability

When a research project is finished, many researchers see the report or articles produced as the only output, and are confident that the library will take care of their preservation. But research in the humanities and beyond are often based on material collected by the researcher, such as ethnographic objects, sound recordings, images, and notes. The scholarly conclusions are then based on such sources. To sustain links from sources to testable conclusions, they have to be stored so that they are accessible to future researchers. But even in museums, this is often done only in a partial manner. Objects may find their way into the collections. But images, recordings and notes are often seen as the researcher's private property and responsibility, and may be lost when her career is ended. Examples of this are hard to document, though, because such decisions are not generally made public.

Conclusion

Sustainability of data in the cultural heritage sector is, as we have seen, not just a technical challenge. The sustainability is eased by the use of open and transparent standards. It is necessary to ensure the existence of well funded permanent organisation like national archives and libraries. Datasets from museums are often not considered to lie within the preservation scope of the existing organisations. Either this has to be changed or the large museums have to act as digital archives of museum data in general. However, the most important measure to ensure sustainability is to increase the awareness of the challenge among curators and scholars. If not, large amounts of irreplaceable research documentation will continue to be lost.

References

DPC(2001): "Digital preservation coalition". http://www.dpconline.org/graphics.

Dunning, Alastair(2001): "Excavating data: Retrieving the Newham archive". http://ahds.ac.uk/creating/case-studies/newham/.

Gausdal, Ranveig Låg (editor) (2006): *Cultural heritage for all — on digitisation, digital preservation and digital dissemination in the archive, library and museum sector.* A report by the Working Group on Digitisation, the Norwegian Digital Library. ABM-Utvikling.

Holmen, Jon; Ore, Christian-Emil and Eide, Øyvind(2004): "Documenting two histories at once: Digging into archaeology". In: Enter the Past. The E-way into the Four Dimensions of Cultural Heritage. BAR, BAR International Series 1227, pp. 221-224

LC(2007): "The library of congress. Digital preservation. http://www.digitalpreservation.gov.

Nygaard, Kristen(1992): "How many choices do we make? How many are difficult? In: *Software Development and Reality Construction*, edited by Floyd C., Züllighoven H., Budde R., and R., Keil-Slawik. Springer-Verlag, Berlin, pp. 52-59.

UNESCO (2003): "Charter on the preservation of digital heritage. Adopted at the 32nd session of the 9/10 general conference of UNESCO" Technical Report, UNESCO. http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf.

Westcott, Keith(2005): Preservation Handbook. *Computer Aided Design (CAD)*. Arts and Humanities Data Service. http://ahds.ac.uk/preservation/cad-preservation-handbook.pdf.

Paper 2: Sustaining Collection Value: Managing Collection/Item Metadata Relationships

Allen H. Renear, Richard Urban, Karen Wickett, Carole L. Palmer, David Dubin
University of Illinois at Urbana-Champaign, USA

Introduction

Collections of texts, images, artefacts, and other cultural objects are usually designed to support particular research and scholarly activities. Toward that end collections themselves, as well as the items in the collections, are carefully developed and described. These descriptions indicate such things as the purpose of the collection, its subject, the method of selection, size, nature of contents, coverage, completeness, representativeness, and a wide range of summary characteristics, such as statistical features. This information enables collections to function not just as aggregates of individual data items but as independent entities that are in some sense more than the sum of their parts, as intended by their creators and curators [Currall et al., 2004,Heaney, 2000,Lagoze et al., 2006, Palmer, 2004]. Collection-level metadata, which represents this information in computer processable form, is thus critical to the distinctive intellectual and cultural role of collections as something more than a set of individual objects.

Unfortunately, collection-level metadata is often unavailable or ignored by retrieval and browsing systems, with a corresponding loss in the ability of users to find, understand, and use items in collections [Lee, 2000, Lee, 2003, Lee, 2005, Wendler, 2004]. Preventing this loss of information is particularly difficult, and particularly important, for "metasearch", where item-level descriptions are retrieved from a number of different collections simultaneously, as is the case in the increasingly distributed search environment [Chistenson and Tennant, 2005, Dempsey, 2005, Digital Library Federation, 2005, Foulonneau et al., 2005, Lagoze et al., 2006, Warner et al., 2007].

The now familiar example of this challenge is the "'on a horse' problem", where a collection with the collection-level subject "Theodore Roosevelt" has a photograph with the item-level annotation "on a horse" [Wendler, 2004] Item-level access across multiple collections (as is provided not only by popular Internet search engines, but also specialized federating systems, such as OAI portals) will not allow the user to effectively use a query with keywords "Roosevelt" and "horse" to find this item, or, if the item is retrieved using item-level metadata alone, to use collection-level information to identify the person on the horse as Roosevelt.

The problem is more complicated and consequential than the example suggests and the lack of a systematic understanding of the nature of the logical relationships between collection-level metadata and item-level metadata is an obstacle to the development of remedies. This understanding is what is required not only to guide the development of context-aware search and exploitation, but to support management and curation policies as well. The problem is also timely: even as recent research continues to confirm the key role that collection context plays in the scholarly use of information resources [Brockman et al., 2001, Palmer, 2004], the Internet has made the context-free searching of multiple collections routine.

In what follows we describe our plans to develop a framework for classifying and formalizing collection-level/item-level metadata relationships. This undertaking is part of a larger project, recently funded by US Institute for Museum and Library Services (IMLS), to develop tools for improved retrieval and exploitation across multiple collections.[1]

Varieties of Collection/Item Metadata Relationships

In some cases the relationship between collection-level metadata and item-level metadata attributes appears similar to non-defeasible inheritance. For instance, consider the Dublin Core Collections Application Profile element *marcrel:OWN*, adapted from the MARC cataloging record standard. It is plausible that within many legal and institutional contexts whoever owns a collection owns each of the items in the collection, and so if a collection has a value for the *marcrel:OWN* attribute then each member of the collection will have the same value for *marcrel:OWN*. (For the purpose of our example it doesn't matter whether or not this is actually true of *marcrel:OWN*, only that some attributes are sometimes used by metadata librarians with an understanding of this sort, while others, such as *dc:identifier*, are not).

In other cases the collection-level/item-level metadata relationship is almost but not quite this simple. Consider the collection-level attribute *myCollection:itemType*, intended to characterize the type of objects in a collection, with values such as "image," "text," "software," etc. (we assume heterogeneous collections).[2] Unlike the preceding case we cannot conclude that if a collection has the value "image" for *myCollection:itemType* then the items in that collection also have the value "image" for that same attribute. This is because an item which is an image is not itself a collection of images and therefore cannot have a non-null value for *myCollection:itemType*.

However, while the rule for propagating the information represented by *myCollection:itemType* from collections to items is not simple propagation of attribute and value, it is nevertheless simple enough: if a collection has a value, say "image," for *myCollection:itemType*, then the items in the collection have the same value, "image" for a corresponding attribute, say, *myItem:type*, which indicates the type of item (cf. the Dublin Core metadata element *dc:type*). The attribute myItem:type has the same domain of values as *myCollection:itemType*, but a different semantics.

When two metadata attributes are related as *myCollection: itemType* and *myItem:type* we might say the first can be v-converted to the other. Roughly: a collection-level attribute **A** v-converts to an item-level attribute **B** if and only if whenever a collection has the value *z* for **A**, every item in the collection has the value *z* for **B**. This is the simplest sort of convertibility — the attribute changes, but the value remains the same. Other sorts of conversion will be more complex. We note that the sort of propagation exemplified by *marcrel:0WN* is a special case of v-convertibility: *marcrel:0WN* v-converts to itself.

This analysis suggests a number of broader issues for collection curators. Obviously the conversion of collection-level metadata to item-level metadata, when possible, can improve discovery and exploitation, especially in item-focused searching across multiple collections. But can we even in the simplest case be confident of conversion without loss of information? For example, it may be that in some cases an "image" value for *myCollection:itemType* conveys more information than the simple fact that each item in the collection has "image" value for *myItem:type*.

Moreover there are important collection-level attributes that both (i) resist any conversion and (ii) clearly result in loss of important information if discarded. Intriguingly these attributes turn out to be carrying information that is very tightly tied to the distinctive role the collection is intended to play in the support of research and scholarship. Obvious examples are metadata indicating that a collection was developed according to some particular method, designed for some particular purpose, used in some way by some person or persons in the past, representative (in some respect) of a domain, had certain summary statistical features, and so on. This is precisely the kind of information that makes a collection valuable to researchers, and if it is lost or inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators.

The DCC/CIMR Project

These issues were initially raised during an IMLS Digital Collections and Content (DCC) project, begun at the University of Illinois at Urbana-Champaign in 2003. That project developed a collection-level metadata schema[3] based on the RSLP[4] and Dublin Core Metadata Initiative (DCMI) and created a collection registry for all the digital collections funded through the Institute of Museum and Library Services National Leadership Grant (NLG) since 1998, with some Library Services and Technology Act (LSTA) funded collections included since 2006[5]. The registry currently contains records for 202 collections. An item-level metadata repository was also developed, which so far has harvested 76 collections using the OAI-PMH protocol[6].

Our research initially focused on overcoming the technical challenges of aggregating large heterogeneous collections of item-level records and gathering collections descriptions from contributors. We conducted studies on how content contributors conceived of the roles of collection descriptions in digital environments [Palmer and Knutson, 2004, Palmer et al., 2006], and conducted preliminary usability work. These studies and related work on the CIC Metadata Portal[7], suggest that while the boundaries around digital collections are often blurry, many features of collections are important for helping users navigate and exploit large federated repositories, and that collection and item-level descriptions should work in concert to benefit certain kinds of user queries [Foulonneau et al., 2005].

In 2007 we received a new three year IMLS grant to continue the development of the registry and to explore how a formal description of collection-level/item-level metadata relationships could help registry users locate and use digital items. This latter activity, CIMR, (Collection/Item Metadata Relationships), consists of three overlapping phases. The first phase is developing a logic-based framework of collection-level/ itemlevel metadata relationships that classifies metadata into varieties of convertibility with associated rules for propagating information between collection and item levels and supporting further inferencing. Next we will conduct empirical studies to see if our conjectured taxonomy matches the understanding and behavior of metadata librarians, metadata specification designers, and registry users. Finally we will design and implement pilot applications using the relationship rules to support searching, browsing, and navigation of the DCC Registry. These applications will include non-convertible and convertible collection-level/item-level metadata relationships.

One outcome of this project will be a proposed specification for a metadata classification code that will allow metadata specification designers to indicate the collectionlevel/item-level metadata relationships intended by their specification. Such a specification will in turn guide metadata librarians in assigning metadata and metadata systems designers in designing systems that can mobilize collection level metadata to provide improved searching, browsing, understanding, and use by end users. We will also draft and make electronically available RDF/OWL bindings for the relationship categories and inference rules.

Preliminary Guidance for Practitioners

A large part of the problem of sustainability is ensuring that information will be valuable, and as valuable as possible, to multiple audiences, for multiple purposes, via multiple tools, and over time. Although we have only just begun this project, already some preliminary general recommendations can be made to the different stakeholders in collection management. Note that tasks such as propagation must be repeated not only as new objects are added or removed but, as new information about objects and collections becomes available.

For metadata standards developers:

1. Metadata standards should explicitly document the relationships between collectionlevel metadata and item-level metadata. Currently we have neither the understanding nor the formal mechanisms for such documentation but they should be available soon.

For systems designers:

2. Information in convertible collection-level metadata should be propagated to items in order to make contextual information fully available to users, especially users working across multiple collections. (This is not a recommendation for how to manage information internally, but for how to represent it to the user; relational tables may remain in normal forms.)

3. Information in item-level metadata should, where appropriate, be propagated to collection level metadata.

4. Information in non-convertible collection-level metadata must, to the fullest extent possible, be made evident and available to users.

For collection managers:

5. Information in non-convertible metadata must be a focus of data curation activities if collections are to retain and improve their usefulness over time.

When formal specifications and tools based on them are in place, relationships between metadata at the collection and item levels will integrated more directly into management and use. In the mean time, attention and sensitivity to the issues we raise here can still improve matters through documentation and policies, and by informing system design.

Acknowledgments

This research is supported by a 2007 IMLS NLG Research & Demonstration grant hosted by the GSliS Center for Informatics Research in Science and Scholarship (CIRSS). Project documentation is available at http://imlsdcc.grainger.uiuc. edu/about.asp#documentation. We have benefited considerably from discussions with other DCC/CIMR project members and with participants in the IMLS DCC Metadata Roundtable, including: Timothy W. Cole, Thomas Dousa, Dave Dubin, Myung-Ja Han, Amy Jackson, Mark Newton , Carole L. Palmer, Sarah L. Shreeves, Michael Twidale, Oksana Zavalina

References

[Brockman et al., 2001] Brockman, W., Neumann, L., Palmer, C. L., and Tidline, T. J. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Digital Library Federation/Council on Library and Information Resources, Washington, D.C.

[Chistenson and Tennant, 2005] Chistenson, H. and Tennant, R. (2005). *Integrating information resources: Principles, technologies, and approaches*. Technical report, California Digital Library.

[Currall et al., 2004] Currall, J., Moss, M., and Stuart, S. (2004). What is a collection? *Archivaria*, 58:131–146.

[Dempsey, 2005] Dempsey, L. (2005). From metasearch to distributed information environments. Lorcan Dempsey's weblog. Published on the World Wide Web at http:// orweblog.oclc.org/archives/000827.html.

[Digital Library Federation, 2005] Digital Library Federation (2005). *The distributed library: OAI for digital library aggregation: OAI scholars advisory panel meeting, June 20–21, Washington, D.C.* Published on the World Wide Web at http:// www.diglib.org/architectures/oai/imls2004/OAISAP05.htm.

[Foulonneau et al., 2005] Foulonneau, M., Cole, T., Habing, T. G., and Shreeves, S. L. (2005). Using collection descriptions to enhance an aggregation of harvested itemlevel metadata. In ACM, editor, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 32–41, New York. ACM/ IEEE-CS, ACM Press.

[Heaney, 2000] Heaney, M. (2000). *An analytical model of collections and their catalogues*. Technical report, UK Office for Library and Information Science.

[Lagoze et al., 2006] Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., and Saylor, J. (2006). Metadata aggregation and automated digital libraries: A retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 230–239, New York. ACM/IEEE-CS, ACM Press.

[Lee, 2000] Lee, H. (2000). What is a collection? *Journal of the American Society for Information Science*, 51(12):1106–1113.

[Lee, 2003] Lee, H. (2003). Information spaces and collections: Implications for organization. *Library & Information Science Research,* 25(4):419–436.

[Lee, 2005] Lee, H. (2005). The concept of collection from the user's perspective. *Library Quarterly*, 75(1):67–85.

[Palmer, 2004] Palmer, C. (2004). *Thematic research collections*, pages 348–365. Blackwell, Oxford.

[Palmer and Knutson, 2004] Palmer, C. L. and Knutson, E. (2004). Metadata practices and implications for federated collections. In *Proceedings of the 67th ASIS&T Annual Meeting (Providence, RI, Nov. 12–17, 2004)*, volume 41, pages 456–462.

[Palmer et al., 2006] Palmer, C. L., Knutson, E., Twidale, M., and Zavalina, O. (2006). Collection definition in federated digital resource development. In *Proceedings of the 69th ASIS&T Annual Meeting (Austin, TX, Nov. 3–8, 2006)*, volume 43.

[Warner et al., 2007] Warner, S., Bakaert, J., Lagoze, C., Lin, X., Payette, S., and Van de Sompel, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*, 7(1-2):35–52.

[Wendler, 2004] Wendler, R. (2004). *The eye of the beholder: Challenges of image description and access at Harvard*., pages 51–56. American Library Association, Chicago.

Notes

1 Information about the IMLS Digital Collections and Content project can be found at: http://imlsdcc.grainger.uiuc.edu/about.asp.

2 In our examples we will use imaginary metadata attributes. The namespace prefix "*myCollection:*" indicates collection-level attributes and the prefix "*myItem:*" indicates item-level attributes. No assumptions should be made about the semantics of these attributes other than what is stipulated for illustration. The current example, *myCollection: itemType*, does intentionally allude to *cld:itemType* in the Dublin Core Collections Application Profile, and "image," "text," "software," are from the DCMI Type Vocabulary; but our use of *myCollection:itemType* differs from *cld:itemType* in entailing that all of the items the collection are of the indicated type.

3 General overview and detailed description of the IMLS DCC collection description scheme are available at: http://imlsdcc.grainger.uiuc.edu/CDschema_overview.asp

4 http://www.ukoln.ac.uk/metadata/rslp/

5 http://www.imls.gov/

6 http://www.openarchives.org/OAI/openarchivesprotocol.html

7 http://cicharvest.grainger.uiuc.edu/

Paper 3: Sustainability of Annotated Resources in Linguistics

Georg Rehm, Andreas Witt, Erhard Hinrichs, Marga Reis

Introduction

In practically all scientific fields the task of ensuring the sustainability of resources, data collections, personal research journals, and databases is an increasingly important topic – linguistics is no exception (Dipper et al., 2006, Trilsbeek and Wittenburg, 2006). We report on ongoing work in a project that is concerned with providing methods, tools, best-practice guidelines, and solutions for *sustainable* linguistic resources. Our overall goal is to make sure that a large and very heterogeneous set of ca. 65 linguistic resources will be accessible, readable, and processible by interested parties such as, for example, other researchers than the ones who originally created said resources, in five, ten, or even 20 years time. In other words, the agency that funded both our project as well as the projects who created the linguistic resources – the German Research Foundation – would like to avoid a situation in which they have to fund yet another project to (re)create a corpus for whose creation they already provided funding in the past, but the "existing" version is no longer available or readable due to a proprietary file format, because it has been locked away in an academic's hidden vault, or the person who developed the annotation format can no longer be asked questions concerning specific details of the custom-built annotation format (Schmidt et al., 2006).

Linguistic Resources: Aspects of Sustainability

There are several text types that linguists work and interact with on a frequent basis, but the most common, by far, are linguistic corpora (Zinsmeister et al., 2007). In addition to rather simple word and sentence collections, empirical sets of grammaticality judgements, and lexical databases, the linguistic resources our sustainability project is primarily confronted with are linguistic corpora that contain either texts or transcribed speech in several languages; they are annotated using several incompatible annotation schemes. We developed XML-based tools to normalise the existing resources into a common approach of representing linguistic data (Wörner et al., 2006, Witt et al., 2007b) and use interconnected OWL ontologies to represent knowledge about the individual annotation schemes used in the original resources (Rehm et al., 2007a).

Currently, the most central aspects of sustainability for linguistic resources are:

- markup languages
- metadata encoding
- legal aspects (Zimmermann and Lehmberg, 2007, Lehmberg et al., 2007a,b, Rehm et al., 2007b,c, Lehmberg et al., 2008),
- querying and search (Rehm et al., 2007a, 2008a, Söhn et al., 2008), and
- best-practice guidelines (see, for example, the general guidelines mentioned by Bird and Simons, 2003).

None of these points are specific to the field of linguistics, the solutions, however, are. This is exemplified by means of two of these aspects.

The use of markup languages for the annotation of linguistic data has been discussed frequently. This topic is also subject to standardisation efforts. A separate ISO Group, ISO TC37 SC4, deals with the standardisation of linguistic annotations.

Our project developed an annotation architecture for linguistic corpora. Today, a linguistic corpus is normally represented by a single XML file. The underlying data structures most often found are either trees or unrestricted graphs. In our approach we transform an original XML file to several XML files, so that each file contains the same textual content. The markup of these files is different. Each file contains annotations which belong to a single annotation layer. A data structure usable to model the result documents is a multi-rooted tree. (Wörner et al., 2006, Witt et al., 2007a,b, Lehmberg and Wörner, 2007).

The specificities of linguistic data also led to activities in the field of metadata encoding and its standardisation. Within our project we developed an approach to handle the complex nature of linguistic metadata (Rehm et al., 2008b) which is based on the metadata encoding scheme described by the TEI. (Burnard and Bauman, 2007). This method of metadata representation splits the metadata into the 5 different levels the primary information belongs to. These levels are: (1) setting, i.e. the situation in which the speech or dialogue took place; (2) raw data, e.g., a book, a piece of paper, an audio or video recording of a conversation etc.; (3) primary data, e.g., transcribed speech, digital texts etc.; (4) annotations, i.e., (linguistic) markup that add information to primary data; and (5) corpus, i.e. a collection of primary data and its annotations.

All of these aspects demonstrate that it is necessary to use field specific as well as generalised methodologies to approach the issue "Sustainability of Linguistic Resources".

References

Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Burnard, L. and Bauman, S., editors (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium.

Dipper, S., Hinrichs, E., Schmidt, T., Wagner, A., and Witt, A. (2006). Sustainability of Linguistic Resources. In Hinrichs, E., Ide, N., Palmer, M., and Pustejovsky, J., editors, *Proceedings of the LREC 2006 Satellite Workshop Merging and Layering Linguistic Information*, pages 48–54, Genoa, Italy.

Lehmberg, T., Chiarcos, C., Hinrichs, E., Rehm, G., and Witt, A. (2007a). Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 164–166, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Lehmberg, T., Chiarcos, C., Rehm, G., and Witt, A. (2007b). Rechtsfragen bei der Nutzung und Weitergabe linguistischer Daten. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen – Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007*, pages 93–102. Gunter Narr, Tübingen.

Lehmberg, T., Rehm, G., Witt, A., and Zimmermann, F. (2008). Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends*. In print.

Lehmberg, T. and Wörner, K. (2007). Annotation Standards. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, Handbücher zur Sprach-und Kommunikationswissenschaft (HSK)*. de Gruyter, Berlin, New York. In press.

Rehm, G., Eckart, R., and Chiarcos, C. (2007a). An OWL-and XQuery-Based Mechanism for the Retrieval of Linguistic Patterns from XML-Corpora. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *International Conference Recent Advances in Natural Language Processing (RANLP 2007)*, pages 510–514, Borovets, Bulgaria.

Rehm, G., Eckart, R., Chiarcos., C., Dellert, J. (2008a). Ontology-Based XQuery'ing of XML-Encoded Language Resources on Multiple Annotation Layer. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Schonefeld, O., Witt, A., Lehmberg, T., Chiarcos, C., Bechara, H., Eishold, F., Evang, K., Leshtanska, M., Savkov, A., and Stark, M. (2008b). The Metadata-Database of a Next Generation Sustainability Web-Platform for Language Resources. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007b). Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 166–170, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Rehm, G., Witt, A., Zinsmeister, H., and Dellert, J. (2007c). Masking Treebanks for the Free Distribution of Linguistic Resources and Other Applications. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT 2007)*, number 1 in Northern European Association for Language Technology Proceedings Series, pages 127–138, Bergen, Norway.

Schmidt, T., Chiarcos, C., Lehmberg, T., Rehm, G., Witt, A., and Hinrichs, E. (2006). Avoiding Data Graveyards: From Heterogeneous Data Collected in Multiple Research Projects to Sustainable Linguistic Resources. In *Proceedings of the E-MELD 2006 Workshop on Digital Language Documentation: Tools and Standards – The State of the Art*, East Lansing, Michigan.

Söhn, J.-P., Zinsmeister, H., and Rehm, G. (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. In *Proceedings of LREC 2008 workshop on Sustainability of Language Resources and Tools for Natural Language Processing*, Marrakech, Morocco.

Trilsbeek, P. and Wittenburg, P. (2006). Archiving Challenges. In Gippert, J., Himmelmann, N. P., and Mosel, U., editors, *Essentials of Language Documentation,* pages 311–335. Mouton de Gruyter, Berlin, New York.

Witt, A., Rehm, G., Lehmberg, T., and Hinrichs, E. (2007a). Mapping Multi-Rooted Trees from a Sustainable Exchange Format to TEI Feature Structures. In *TEI@20: 20 Years of Supporting the Digital Humanities*. The 20th Anniversary Text Encoding Initiative Consortium Members' Meeting, University of Maryland, College Park.

Witt, A., Schonefeld, O., Rehm, G., Khoo, J., and Evang, K. (2007b). On the Lossless Transformation of Single-File, Multi-Layer Annotations into Multi-Rooted Trees. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2007*, Montréal, Canada.

Wörner, K., Witt, A., Rehm, G., and Dipper, S. (2006). Modelling Linguistic Data Structures. In Usdin, B. T., editor, *Proceedings of Extreme Markup Languages 2006*, Montréal, Canada.

Zimmermann, F. and Lehmberg, T. (2007). Language Corpora – Copyright – Data Protection: The Legal Point of View. In Schmidt, S., Siemens, R., Kumar, A., and Unsworth, J., editors, *Digital Humanities 2007*, pages 162–164, Urbana-Champaign, IL, USA. ACH, ALLC, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign.

Zinsmeister, H., Kübler, S., Hinrichs, E., and Witt, A. (2008). Linguistically Annotated Corpora: Quality Assurance, Reusability and Sustainability. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics, HSK*. de Gruyter, Berlin etc. In print.

ALLC SESSION: e-Science: New collaborations between information technology and the humanities

Speakers

David Robey
d.j.b.robey@reading.ac.uk
Arts and Humanities Research Council, UK

Stuart Dunn
stuart.dunn@kcl.ac.uk
King's College London, UK

Laszlo Hunyadi
hunyadi@ling.arts.unideb.hu
University of Debrecen, Hungary

Dino Buzzetti
buzzetti@philo.unibo.it
University of Bologna, Italy

e-Science in the UK and elsewhere stands for a broad agenda as important for the humanities as it is for the natural sciences: to extend the application of advanced information technologies to develop new kinds of research across the whole range of academic disciplines, particularly through the use of internet-based resources. The aim of this session is to introduce some recent UK developments in this area, side by side with related developments in other parts of Europe.

• David Robey: Introduction

• Stuart Dunn: e-Science developments in the UK: temporal mapping and location-aware web technologies for the humanities

• Laszlo Hunyadi: Virtual Research Organizations for the Humanities in Europe: technological challenges, needs and opportunities

• Dino Buzzetti: Interfacing Biological and Textual studies in an e-Science Environment