Toward Best Practice for Language Resource Conversion

EMELD 2003 Working Group on Resource Conversion

Working Group

- Baden Hughes, Chilin Shih (co-chairs)
- Helen Aristar-Dry, Steven Bird, Reinhard Hiss, Will Lewis, Barbara Need, Steven Weinberger

Objectives

- Consider the methodology for and make recommendations about the conversion of legacy (possibly non-digital) language resources into enduring BP formats
- Examine ongoing conversion processes and identify issues in the conversion of digital language resources in working contexts

Methodology

- Focus on high level principles which pervade general language resource conversion problems rather than format-specific resource conversion issues
- Acceptance that appropriate technical expertise probably already exists "somewhere" but needs to be adapted to the EMELD context

Subject Matter

- Content and Structure
  - Metadata
  - Text
  - Audio
  - Video
  - Still Images
- Physical Media
- Hardware / Software

Core Values

- Bird & Simons (2003) "Seven Dimensions …": content, format, discovery and preservation
- Motivation to ensure persistence and longevity of archive quality digital objects

Principles …1

- Ignorance is not bliss !
- Not every user needs to be a technical expert, but should be assisted their context and functional requirements and to access sufficient information to make an informed choice
- Conversion issues will affect institutions and individuals at many levels – particularly in terms of resources available to address issues

Principles …2

- Conversion and Archiving
  - The best available copy should be archived according to BP
  - Format neutrality in respect to use involves effort but is essential to ensure long term viability
  - Archiving practice will imply resource conversion for preservation purposes
  - Consistency in conversion methodology is inherently better than random variation

Principles …3

- Conversion and Re-Use
  - Requirements for re-use vary between agents and purposes
  - Inherent in most (all?) conversion processes is some degree of information loss, thus the absolute minimum number of format conversions should be undertaken
  - Where possible, converted materials should include information about their digital lineage
  - Additional information pertaining to the language resource may be located separately from the resource itself and needs to be preserved

A Pragmatic Approach to BP .. 1

- The lineage of digital language resources may have included processes which are less than optimal practices
- BP may not realistically be achievable in all contexts (constraints such as time, money, equipment, expertise, inclination …)
- Some practices have inherently higher potential to cause conversion and archiving issues
- Significant incentives need to be offered to induce change in language data management practices towards BP – would you prefer to choose BP or be forced to adopt BP when you lose data ?

A Pragmatic Approach to BP .. 2

- Software choice will impact on the longevity of language resource data.
- Ideological debates about software development methodologies is often misleading when considering longevity and preservation
- Absolute ranking of practice on a scale of worst to best is not transparent (context sensitive, moving target …)

Ongoing Work Items …1

- Identify and review core documents on BP formats, including accessible recommendations for different audiences
- Identify and review software tools which enable conversion according to BP principles (this is not necessarily a democratic system!)
- Develop accessible case studies of typical language resource conversion problems, critique them and provide advice on how to achieve BP in these contexts

Ongoing Work Items … 2

- Examine how physical media choices can affect the retention or loss of information and implications for the language resource conversion process
- Promulgate resource conversion as a pervasive issue to be considered by many other BP contexts

Observations Relevant to Other Working Groups

- Resource Archiving
  - Good archiving practice will consider resource conversion as a fundamental issue
  - Infrastructural constraints may significantly increase the risk of information loss
- Resource Creation
  - BP at the data collection point reduces the risk of information loss in any conversion process
  - Conversion implications need to be considered when selecting an appropriate tool for the data and functionality types required

Observations Relevant to EMELD

- EMELD needs to consider the longevity and persistency implications for ongoing archiving functions particularly in reference to the "long term" – this may include adequate financial resourcing

Logistical Recommendations

- Creation of Communities of Expertise within EMELD framework to advise on working group topics (cf. Ask-A-Linguist) including experts from outside linguistics
- Creation of Working Groups email lists for ongoing work in these areas
- User reviews and solutions section for tools and processes within the EMELD School site