

# Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



## Toward the Interoperability of Language Resources

Nick Thieberger

Department of Linguistics & Applied Linguistics, The University of Melbourne

### 1. What are the greatest barriers to interoperability?

Lack of accepted standards for data formats. Focus on linguistic analysis at the expense of data means that few linguists consider interoperability to be a problem. Lack of training for practitioners in use of tools and in understanding reasons for producing well-formed data. Reluctance of many established linguists to engage with the possibilities offered by new technologies.

### 2. What could this workshop do to best promote interoperability?

Identify standards and propose them to OLAC. Identify where standards are lacking and establish ways of building them.

### 3. What sets of tools or facilities have you used that are currently interoperable? What are the benefits and drawbacks of these and/or similar tool suites?

The metadata for our collection (at PARADISEC) exports to a form conformant to the OLAC metadata standard and has been a good model of how well structured data can be addressed by independent services based on its conformance to published standards. With the benefit of some experience my understanding of data structure now is that any tool that allows you to keep each structural element identified in some way can be useful. Familiarity with the tool is a big incentive to keep using it, so adapting known tools to produce good data is a viable approach.

### 4. If you have been involved in tool development, what are the primary challenges involved in designing interoperable tools?

I am working on EOPAS, an XML schema for the online presentation of interlinear text linked to media. We selected the outputs of Transcriber, Elan and Toolbox as the inputs to EOPAS but had to constrain the form of the data in these tools to allow it to be converted to the common schema. While the data in each of these is typically reusable in that the underlying form of the data is a text file, the unconstrained nature of data in each tool means that it is not easily made interoperable.



Establishing standards-conformant templates for these tools is one way to enforce interoperability.

5. If your work involves a range of non-interoperable tools, what solutions or work-arounds have you found?

It is the nature of the data that should be primary, and that can mean using various tools. The process of conversion between tools (the workflow) means I use regular expressions to convert structures as required.

6. Do you agree that interoperable tools will produce interoperable documentation, and this in turn will facilitate the development of Internet services and digital archives. Or does this claim require qualification or explication?

I'm not sure what an interoperable tool is, but have found that good tools in poor hands can have worse results than poor tools in good hands, where the good hands are those that understand issues of creating reasonably normal data structures.

Otherwise I agree with this claim, but recognise that most linguists simply do not understand the issues and that we have to provide training, standards-based tools and examples of workflows resulting in ideal outputs. An example of a tool that is gaining popularity is Toolbox and its success is partly due to the ease with which data can be exported via MDF and LexiquePro to provide paper and web-based outputs.

We really do need published standards against which data can regularly be validated so that it can be accessed in useful ways, for example, interlinear text being addressed by services that know what interlinear text is so that it is not simply treated as a text file.

