

# Data Interoperability: Local Control versus Global Integration

Scott Farrar

## 1 Issue statement

It is by now widely accepted that using a common markup language, such as XML, and a common character encoding, such as Unicode, can greatly facilitate the sharing of data among diverse applications by achieving ‘syntactic interoperability’. However, the ability to use diverse data sets—even within the same application—is severely limited by ‘content interoperability’. That is, even though two data sets may be syntactically similar, the application has no way to interpret the meanings of individual data elements. Consider, for example, two data sets that both refer to the same fundamental concept for which one data set uses a type called SYNTACTICCATEGORY while the other uses a type called PARTOFSPEECH. Applications would have no way of determining that the two types were in fact compatible. One solution is simply to require that data be conformant to some standard not only syntactically, but also in terms of content. This approach, however, introduces a top-down management paradigm that many individual data providers would not be ready to accept. It is proposed here that the issue of content interoperability poses the biggest challenge to a cyberinfrastructure for linguistics. Thus, the general issue to be discussed here is how to create tools that facilitate global content interoperability, while allowing the individual linguist the ability to maintain local control over individual data sets. Related to this are three specific talking points:

First, there is the issue of how to define the notion of interoperability, specifically for linguistic data. From a data-centric point of view, we can preliminarily define data interoperability as the ability of an application to be able to manipulate various types of linguistic data. In one such scenario, an application is used to search over two different kinds of data, for instance, over both lexical and textual data. The application should be able to give results about language *per se*, results that factor out any particular data structure used for containing language elements. Included in this discussion point are an enumeration of what the common linguistic data types are and how their common factors can be identified.

Another issue concerns the issue of data migration. While convincing linguists to adhere to certain standards may not be feasible, it is more realistic to imagine that diverse data—even unstructured ‘legacy’ data—will be migrated to an interoperable form, a kind of interlanguage for the cyberinfrastructure. That is, specific tools and resources specifically designed to perform the migration process should be discussed. The issue of migration relates directly to the issue of local control. How often should locally maintained data be transformed, and by whom? What sort of bookkeeping mechanisms should be designed to manage the periodic migration?

Finally, there is the issue of how to utilize current standards and technology to achieve a content interoperability. The first set of (emerging) standards includes very structured markup languages such as the Resource Description Framework (RDF) and the OWL Web Ontology Language. Collectively, the community of practice that seeks to utilize these standards for content interoperability is referred to as the ‘Semantic Web’ initiative. It is worth exploring whether the development of a Semantic Web for linguistics a viable initiative for the field? How does this relate to a cyberinfrastructure? Second, there are the various XML-related languages such as the Extensible Stylesheet Language (XSL), XML Query, XPath, and XPointer. And third, there are several Web-related protocols such as the Simple Object Access Protocol (SOAP). What role do such protocols play in developing the cyberinfrastructure?