

Digitization of legacy literature is currently a big issue, e.g., Biodiversity Heritage Library, AMNH digital library and antbase.org. In order to preserve the structure of the documents, e.g., paragraphs, they can be marked up with XML. Additional XML markup may encode the logical structure of systematics publications such as the description of taxa. TaxonX and TaXMLit are two candidate schemas for this purpose (see related abstracts). These schemas build upon the fact that taxonomic publications are highly structured and standardized, each of their elements related to a particular taxon. The main structural elements include the description of the species, nomenclature, distribution, materials examined, tools for identification, phylogenies, illustrations, and bibliographic references. Such marked-up documents allow more detailed search and text mining than provided by other digital library projects for taxonomic literature. This approach does however call for specific tools to automatically create this fine-grained markup.

GoldenGATE is a dedicated XML editor to encode taxonomic publications using taxonomy-specific schemas. Manually creating such detailed markup, which can reach down to the sentence level and below is cumbersome, time-consuming and therefore expensive. Automation is desirable wherever possible. Bio-NLP has lately developed algorithms like TaxonGrab (Koning et al, *TaxonGrab: Extracting Taxonomic Names from Text*, Biodiversity Informatics, 2005) and FAT (Sautter et al, *A Combined Approach to Find all Taxon Names (FAT) in Legacy Biosystematics Literature*, submitted for publication), which identify taxonomic names in texts. NLP tools for the recognition of (collecting) locations have existed since the late 1990s, even though they are not specifically intended for bio-informatics. Both types of tools can significantly reduce the manual effort of markup. However, they do not achieve 100% accuracy, implying the need for manual corrections. Manual steps usually rely on XML editors like XMLSpy or Oxygen. But the purpose of these editors is handling existing XML data rather than creating XML documents from plain text. With these editors, marking up a document means applying NLP tools first, and then doing the rest of the markup in the editor manually, including the correction of NLP errors. The requirement to go back and forth between NLP tools and an XML editor induces further effort. Our GoldenGATE editor is designed to tightly integrate the NLP application and provide as much automation as possible to the manual markup. GoldenGATE integrates existing NLP tools through a slim programming interface. Implementations of this interface currently exist for FAT and a location extractor. Manually inserting XML tags works by selecting the tag content (the selection automatically extends to word boundaries) and creating the tag by simply selecting the XML element name. Further features of GoldenGATE comprise sequencing of automated editing steps to Pipelines, automatically processing a set of files, basic NLP (gazetteers, regular expressions), and basic markup transformation and filtering. Preliminary experiments show that GoldenGATE incurs significant performance gains over conventional XML editors.

[The GoldenGATE is available at http://idaho.ipd.uni-karlsruhe.de/GoldenGATE/.](http://idaho.ipd.uni-karlsruhe.de/GoldenGATE/)