

Wouter Alink (MSc)  
Department of Digital Forensics and Biometry  
Netherlands Forensic Institute

## **A comparison of markup problems found in multilevel text-processing and digital forensics**

This presentation will give an overview of the problems and solutions found both in the Natural Language Processing community as well as in the digital forensics community with respect to the use of stand-off annotation and multi-dimensional markup in general [2].

Natural Language Processing(NLP) can be described as the computational modeling and processing of human language, often towards the understanding of language. One of the areas of NLP is the syntactical analysis and annotation of textual data and the querying thereof [1]. Digital forensics, amongst others, involves the analysis of digital evidence derived from digital sources [3]. Recently this analysis process has been shifting towards the automated extraction of features from digital evidence, and storing those in a database, for later retrieval by a forensic investigator [4].

The automated extraction of the features can in many aspects be compared to the process of annotating textual content and the result thereof. Both deal with large amounts of data, and vast amounts of generated annotations. The majority of these generated annotations will never (or hardly) be used during querying. Furthermore, the use of stand-off annotation in both areas has been proposed as a solution for dealing with the multi-dimensional markup [5, 6]. An important difference between the two areas is that NLP has been a scientific research area for a few decades, while the digital forensics community is fairly new, and has originated from a very practical need. Can digital forensics and NLP learn from each other?

Keywords: stand-off annotation, multi-dimensional markup, text-processing, querying, digital forensics, XML

### References:

- [1] P. Ogilvie. 2004. Retrieval Using Structure for Question Answering. In The First Twente Data Management Workshop (TDM'04), pages 15–23
- [2] H.S. Thompson and D. McKelvie. Hyperlink semantics for standoff markup of read-only documents. In Proceedings of SGML Europe '97, Barcelona, Spain, May 1997
- [3] Digital Forensic Research Workshop, <http://www.dfrws.org/>
- [4] Tye Stallard and Karl Levitt, Automated Analysis for Digital Forensic Science: Semantic Integrity Checking, 2003
- [5] Workshop Multi-dimensional Markup in Natural Language Processing (NLPXML-2006), EACL 2006
- [6] W. Alink, R.A.F. Bhoedjang, P.A. Boncz, A.P. de Vries, XIRAF - XML-based indexing and querying for digital forensics. In proceedings of DFRWS 2006.