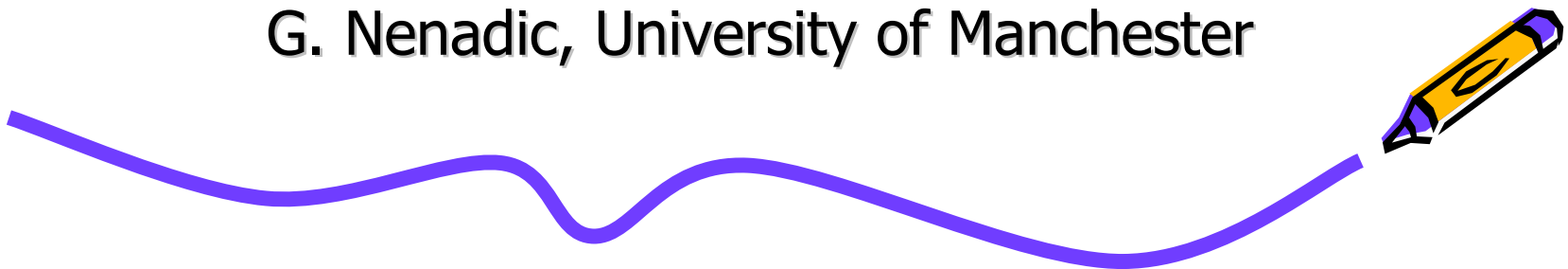
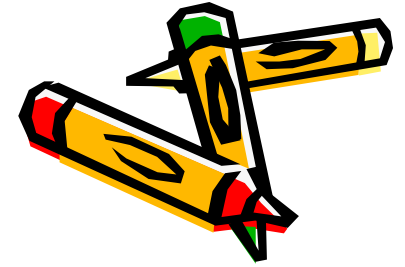


Towards standardisation of named-entity annotations in the life science literature

D. Rebholz-Schuhmann, EBI

G. Nenadic, University of Manchester

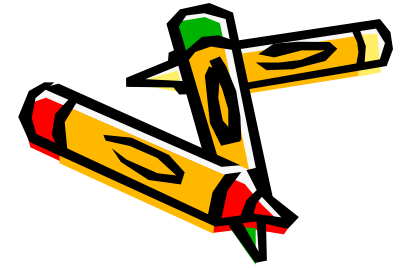




Overview

- Sharing of annotated documents is still not happening
=> make another effort to solve this
- Beneficiaries
- IeXML proposal: now a working example (?)
- Plan for future activities

The main idea: How to represent semantics in text (as part of text)



Text miners' view:

Term => Named Entity => Semantic type

Ontologists' view:

Concept => Label => Evidence in Text

Bioinformaticians/"Users"/Data Miners' view:

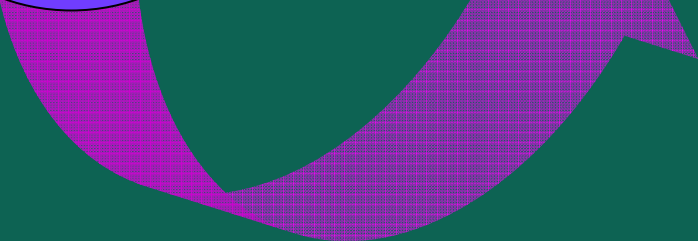
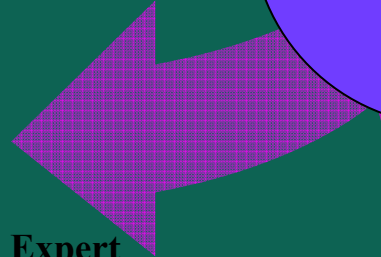
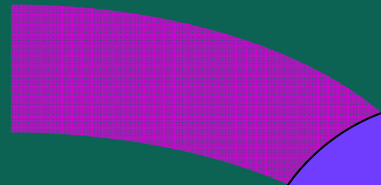
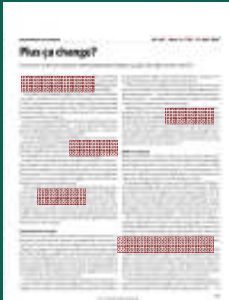
Text passage <=> Semantics

The main idea (2):



- Named entity annotation is still a challenge:
 - Resources for NEs are not standardized => standard semantics
 - Extraction methods are available, but are difficult to integrate in third party solutions
 - No minimum standard suitable for large-scale corpora (biomedical literature) available
- Additional constraints:
 - Avoid any costs (manual post-processing, SW maintenance)
 - Has to harmonize with a number of document formats (publishers)

leXML?

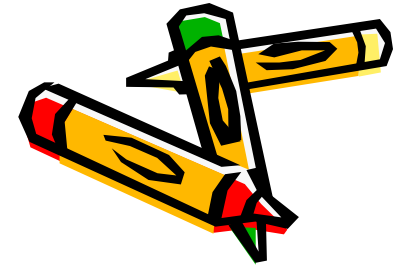


Group 2

Expert

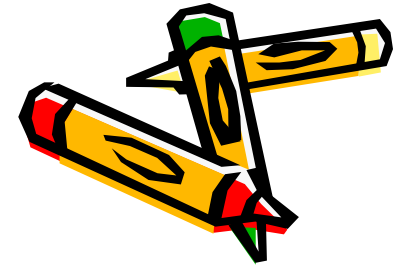


Group 1



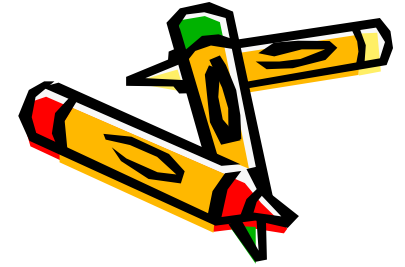
Beneficiaries

- **users:** facilitate semantics-based browsing, visualisation, integration without need to do TM
- **text mining community:** compare and improve the state of the art in NER, and motivate progress in other text mining tasks; facilitate interoperability of results
- **publishers and industry:** provide an added value to their products, and thus facilitate data sharing, availability and interoperability.



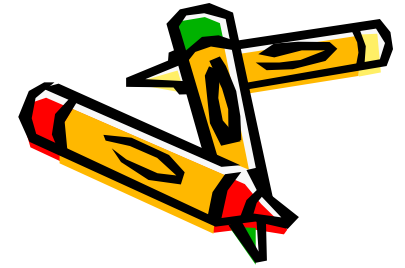
What benefits for TMiners

1. Provide annotated document repositories
 - both results of automated text mining and publishers efforts for semantic annotations
 - document exchange
2. Many (?) other bio-data is in standardised formats, why not textual/NE data?
3. Many in-house formats are not productive
 - processing overheads
 - not easy to compare
 - focus on other tasks



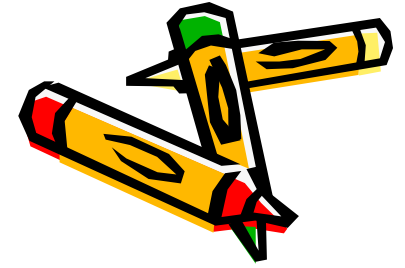
Working towards

1. Recommendations for a common syntax to embody entity mentions in publishers' document formats
2. Provision of a common way to reference semantic types



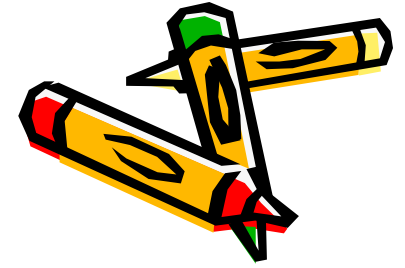
1. Syntactic level

- identification of a **minimal** set of NE tags and features to be included in publishers' formats
- representation of nestedness, ambiguities and multiple annotations (e.g. annotations from different groups/services);
- both inline and stand-off (automatic conversion available)



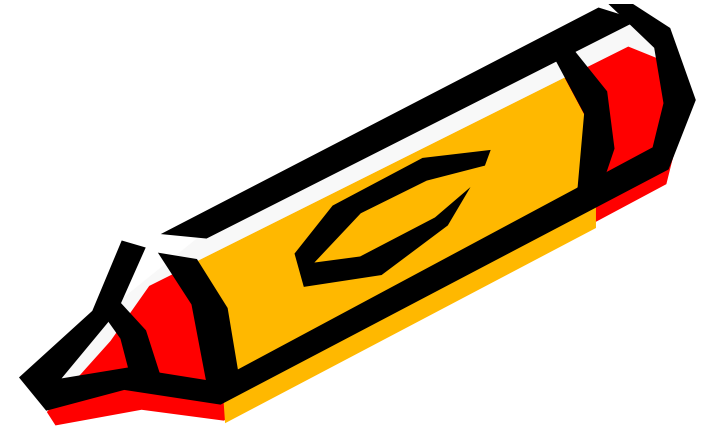
2. Semantic level

- integration of a basic semantic type system into NE/document formats,
- provisions for references/pointers to external type systems
 - e.g. existing ontologies or purposely-built type systems
 - plus, mapping to a top level ontology (?) (UMLS semantic types, BioTOP)

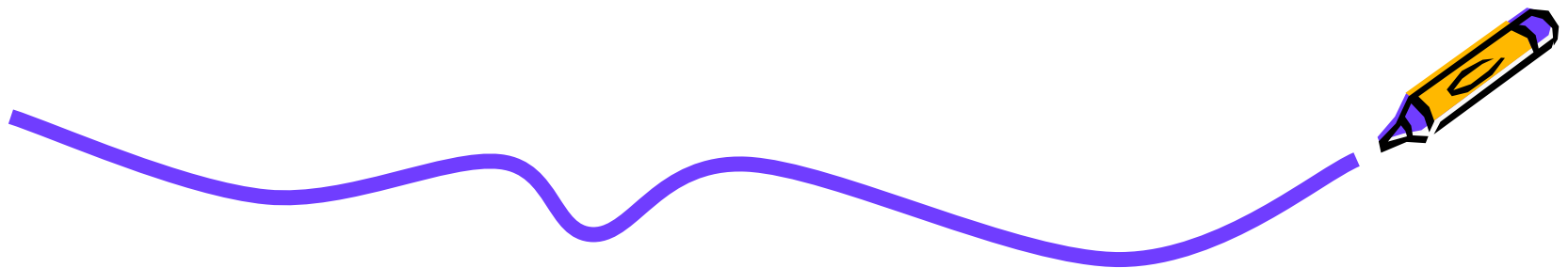


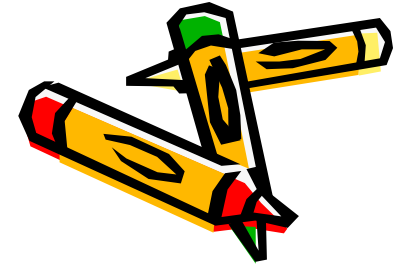
Further benefits/aims

- First step in “semantic enrichment” of literature and linking to other bio-resources
 - define and describe tags in documents
 - articles become part of the Semantic Web
- Support tool interoperability by common formats
 - no need for shim libraries
- Industry-wide standards for NEs
 - still, no need to standardise publishers’ formats



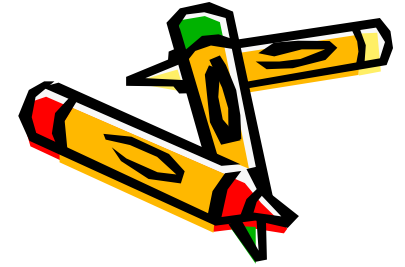
IeXML – a working example





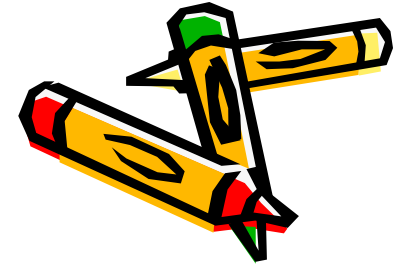
History

- Discussions from 2005
 - SMBM 2005, ISMB 2006
 - Meetings with UK bio-text mining community
 - Comments on BioNLP mailing list, K. Cohen
- Outcomes / common understanding
 - interoperability is needed
 - NEs form the center ground to reach agreements
 - more work is needed



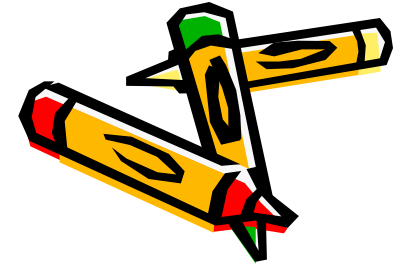
IeXML: aim

- an XML-based annotation scheme with a ***minimal*** tag set together with ***basic*** attributes that each text mining module has to “understand” (inline and standoff)
- simple and scalable (adding proprietary attributes)
- a sample suite of compliant modules (e.g. readers-consumers, taggers, etc.)



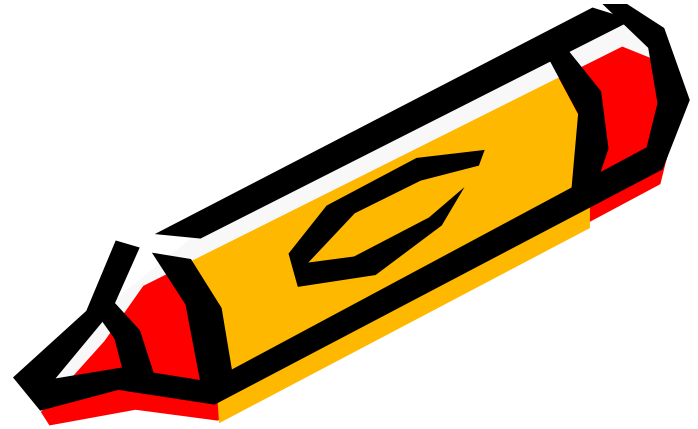
IeXML document

- *text* elements identify stretches in documents that are of interest/target to text mining modules
- entities are annotated using *e* elements
 - one or more tokens or other entities
 - attributes: *sem* (required) and *c* (optional)
- this is a minimal set of annotations

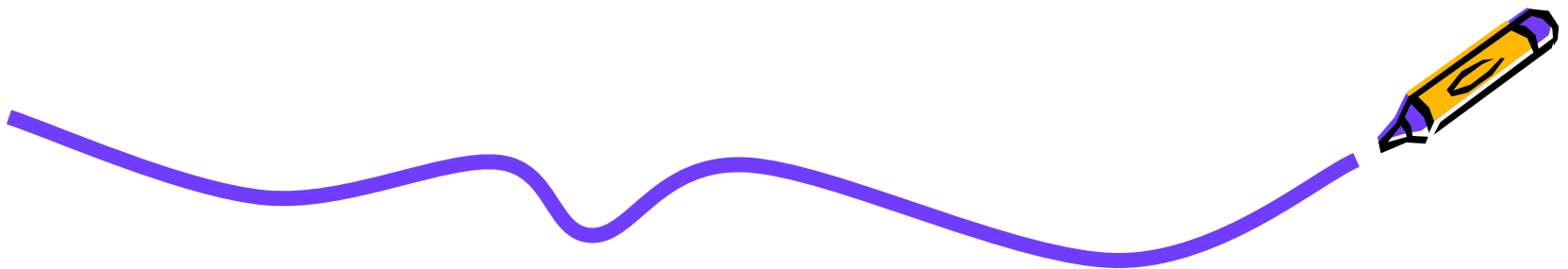


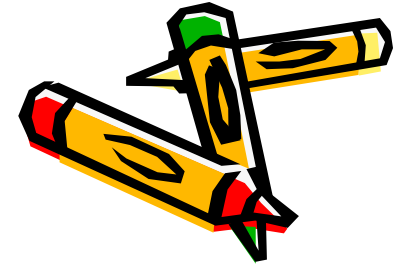
Entity elements – details

- values of **sem**-attribute linking to external resources
- e-elements can be nested
- **c**-attributes is for POS tag
- Other attributes can be used which need not necessarily be in conflict with the basic format
- Further specification needed for ambiguity, semantic types, overlapping NEs



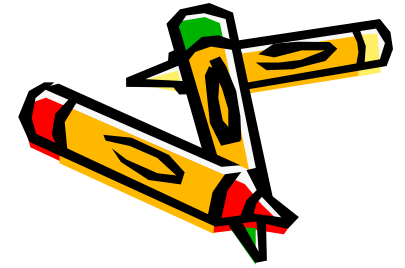
Next steps





Next steps

- Mailing list + Wiki available for collaboration
- Working group anticipated to
 - prepare/review the recommendations
 - settle the document format
- Implement a repository (e.g. BMC, UK PMC) in a new format, with contributions from different groups
- Feedback on annotation variation



Let's share documents with
annotated NEs