# Data models and the (blind ?) query of lexical resources

Laurent Romary

# Data models and the (blind ?) query of lexical resources

*Laurent Romary, Inria & HUB (Saclay, Berlin)*

Note: this abstract is conceived as an outline of both a possible presentation at the workshop and of a possible future paper that could follow it. It thus contains various lacunae that should be filled up step by step.

I would like to address in this note the relation between the query process and the actual model that the data to be queried is based upon, in the context where this model is only partially implemented, as is often the case in the XML world. My use case will be that of the query of a pool of lexical databases, which, I assume here, are all encoded according to the TEI guidelines.

Having in mind various background works on the issue of typing and querying (see seminal works by Al-Khalifa et alii, 2002 and Chamberlin et alii, 2001), I would like to limit my scope here to a) the identification of possible quality constraints we would want to apply to such lexical databases and b) identify work-arounds for dealing with silent data during a query process.

## The lexical data landscape

Lexical resources occur in a wide variety of formats and for a wide variety of application. From basic full-form lexica for NLP to encyclopaedic dictionaries intended for human consumption, the corresponding data formats are constrained by many different factors mainly related to available software (e.g. Shoebox) or scholarly traditions (the Multext format for full-form lexica). From a conceptual point of view, lexical resources can be broadly classified as belonging to one of the two core models: semasiological (word to sense; as implemented in traditional dictionaries) or onomasiological (concept to term; as is the case for most terminological databases). Such a classification is covered in the current standardisation landscape by two complementary ISO standards providing a specification platform for lexical models: ISO 16642 (TMF; see Romary, 2001) for onomasiological models and ISO 24613 (LMF). Furthermore such standards find their natural serialisation in ISO 30042 (TBX) for TMF and, as we argued in (Romary, 2013), by the TEI guidelines. In this note, we will only consider semasiological models, but a similar analysis could be carried out for onomasiological ones, even if these present less complexity, given the rather rigid nature of the TMF model.

## Types of queries

It is of course impossible, without a real large-scale user survey (but maybe our workshop could be the opportunity to carry out something on the TEI-list) to cover exhaustively the query use cases that current dictionary projects or users may have in mind. Still, we can identify a few reference query templates that could serve as a plausible background for our discussion.

Indeed, considering the core components of the semasiological model, as reflected in the LMF core package, namely the entry, the form and the sense we can identify a few classes of queries:

- Retrieval of a specific entry considering constraint on the form, as typically the case for a basic POS tagging task (token to word-form mapping to take up the ISO 24611 (MAF) terminology)
- Retrieval of a sense from an entry given additional constraints
- Search for all entries having some specific form, grammatical or semantic properties, for instance the retrieval of all transitive verbs
- Extraction of all (or part of all) occurrences of a certain descriptor in a group of lexical entries , for instance all translated examples
- …

## Select-from-where and DB models

The literature in database management system is replete with works on the link between the database models and the corresponding query structures (in the select-from-where paradigm) that they authorise. Recent works () have provided some clues concerning the role of patterns (e.g. pairs, sequences, etc.) as typing mechanisms and thus selection keys on semi-structured data. Still very few works have tackled the issue of data with low conformance to a reference model for which very little prediction can be made as to affordable queries.

## Dealing with unpredictable (TEI) data

If we now consider the TEI guidelines as a data model for the representation of lexical data, we can just observe that the variety of constructs it allows makes it more or less impossible to predict the actual micro-structures that a specific lexical database will have. A quick overview of the current state of development of the guidelines offers indeed several levels of variability:

- the entry point to a lexical entry may be based on various elements, <entry>, <superEntry>, <entryFree>, <dictScrap>, all having their own organisational constraint (if at all in the case of <dictScrap>);
- most of the TEI descriptors for dictionary (e.g. <pos>) can potentially appear almost anywhere in a dictionary representation, even if recent works on the TEI guidelines have reduced some of these possibilities (e.g. forcing the use of <gramGrp> to group together grammatical descriptors)
- there is in general no way to ensure that from a given entry to another the same encoding principles will be applied. For instance an entry with one single sense may be either represented with a <sense> element grouping all semantic descriptors, or just left them occur as children of the <entry> element

## Conformance to models – the LMF case

There is thus no choice but complement the TEI guidelines with additional constraints that may ensure some minimal predictability criteria for TEI encoded lexical databases. As systematically recommended recently (Budin & Moerth, 2012; Romary & Wegstein, 2012) and in a way using a further argument to that which led to identify

the TEI guidelines, we should establish a set of LMF based quality standards for TEI lexica that could improve their integration in wider pools of data. This is in particular made necessary by the definition of trans-national initiatives such as DARIAH or CLARIN to pool together networks of digital resources in the humanities.

The compliance with LMF ensures that one can express abstract queries based on combinations of components from the LMF data models and associated data categories. Such queries would then be mapped onto concrete XML queries applicable to the corresponding TEI encoded data.

## Consequences for TEI representations

The LMF conformance of TEI encoded data has some strong consequences on the constraints imposed on the default TEI model.

- Imposes the use of specific components (crystals) in the TEI structure (entry, form, gramGrp, sense in specific configurations)
- Forbids the occurrence of "orphan" descriptors in dictionary entries, namely outside the above-mentioned elements in the dictionary crystals

A priority should be thus set in implementing check-out procedures that warrant that lexical database present a minimal signature according to this constraints. Once a database is conformant to such constraints, we know it can be queried through this abstract model mechanism outlined above.

## Identifying silent data

A complementary consequence to the strategy outlined in this note is that it is possible to contemplate ways of retrieving silent data, that is lexical entries which, for a given query have not produced any results because one of the components or data category mentioned in the query is missing in the representation. A typical case could be to inform the user that a full-form lexicon has produced no hits related to a query on the sense component.

## Going further — data seal of approval for lexical data

The long-term objective of the line of thoughts expressed is basically two fold:

- Constraining lexical database to conform to a general model in order to be incorporable within a larger interoperable pool of data
- Identifying sub-classes of lexical model that may respond to specific types of queries

## References

ISO 16642:2003 Computer applications in terminology -- Terminological markup framework (TMF)

ISO 24610-1:2006 Language resource management -- Feature structures -- Part 1: Feature structure representation

ISO 24611:2012 Language resource management -- Morpho-syntactic annotation framework (MAF)

ISO 24613:2008 Language resource management - Lexical markup framework (LMF)

ISO 30042:2008 Systems to manage terminology, knowledge and content -- TermBase eXchange (TBX)

S. Al-Khalifa, H.V. Jagadish, N. Koudas, J.M. Patel, D. Srivastava, W. Yuqing (2002) "Structural joins: a primitive for efficient XML query pattern matching," 18th International Conference on Data Engineering, 2002. pp.141,152, 2002 — doi: 10.1109/ICDE.2002.994704

Budin Gerhard , Stefan Majewski and Karlheinz Mörth (2012) "Creating Lexical Resources in TEI P5", jTEI, Issue 3.

Chamberlin D. , J. Robie, Florescu, D. (2001) "Quilt: An XML Query Language for Heterogeneous Data Sources" pp. 1-25 in G. Goos, J. Hartmanis, J. Leeuwen, D. Suciu, G. Vossen, (Eds.) *The World Wide Web and Databases,* Springer. — doi: 10.1007/3-540-45271-0_1

Ide N. and J. Véronis, (1995). Encoding dictionaries. In Ide, N., Veronis, J. (Eds.) *The Text Encoding Initiative: Background and Context.* Dordrecht: Kluwer Academic Publishers, 167-80.

Lee Kiyong , Lou Burnard, Laurent Romary, Eric De La Clergerie , Thierry Declerck, Syd Bauman, Harry Bunt, Lionel Clement, Tomaz Erjavec, Azim Roussanaly, Claude Roux (2004) "Towards an international standard on feature structures representation" 4th International Conference on Language Resources and Evaluation - LREC'04 373-376 — http://hal.inria.fr/inria-00099855

Romary L. (2001) An abstract model for the representation of multilingual terminological data: TMF - Terminological Markup Framework. *TAMA 2001*, Antwerp, Belgium. — http://hal.inria.fr/inria-00100405

Romary L. and W. Wegstein (2012), "Consistent modelling of heterogeneous lexical structures", *Journal of the Text Encoding Initiative*, Issue 3 | November 2012, Online since 15 October 2012, connection on 09 January 2013. URL : http://jtei.revues.org/540 ; DOI : 10.4000/jtei.540 — http://hal.inria.fr/hal-00704511

Romary L. (2013) "TEI and LMF crosswalks". In Stefan Gradmann and Felix Sasaki (Eds.). Digital Humanities: Wissenschaft vom Verstehen, Humboldt Universität zu Berlin, 2013 — http://hal.inria.fr/hal-00762664