

# Silfide: A System for Open Access and Distributed Delivery of TEI Encoded Documents

Laurent Romary, Patrice Bonhomme, Florence Bruneseaux, Jean-Marie Pierrel

Laboratoire Loria

B.P. 239, F-54506 Vandœuvre Lès Nancy

{romary,bonhomme,brunesea,jmp}@loria.fr

## 1.Introduction

In recent years, interest in studies based on computerized linguistic resources has revived. Such studies may be in linguistics, literature, or history, as often as in the field of computer science and computational linguistics. Several recent publications (e.g. Aarts et al. 92, TAL-95, IJCL-96) as well as diverse methods and targets used by researchers in these fields testify to this resurgence. The renewal of interest in these methods raises essential questions regarding the status, as well as the maintenance of, digital data. It currently seems unfeasible to repeat *ad infinitum* the traditional working cycles on data. For example, in a given research project, the necessary data is defined, collected and *ad hoc* tools are rapidly constructed to extract the relevant information for the current study. Finally, when the work is complete and the results are published, the data is shelved under a more or less identified form and above all, is only known to the researchers who took part in the research project. In most cases, these data are not reusable for any other project, either because compatible computer tools no longer exist for display formats that have not been documented, or because it would be too expensive to convert these data to make them compatible with new tools defined for new research. Until now, this non-compatibility has made it impossible to consider a flexible and modular use of the data in large, well-known text collections since their individual formats have never been associated with the tools available within the academic community.

In this paper we would like to expand upon the complex issue of data reusability raised by the problems that we describe above.

First, what are the linguistic resources that have to be represented? The case of textual data seems straightforward at first, because they imply a low degree of structure. But even in the simplest modes of representation (untagged texts) we need to add a minimum of documentation (e.g. origins, contents of corresponding texts). It is also essential (M.-P. Pery-Woodley, 95) to collect textual data from complete and identifiable texts so as to master all parameters (gender, structure) which might be used in later studies. Hence each "text" must be considered as an individual entity, rather than considering "texts" *en masse*.

Other linguistic resources tend naturally to be more structured and thus require more processing to make them available to a large community. These include lexical resources which will either take the form of a computerized dictionary (for human use) or that of a lexical database (e.g. as a basic input of a part of a speech tagger). In this last case it is essential to normalize the structure completely, so that the linguistic resources can be integrated into different software platforms. In the same way, there now exist a great number of dialogue corpora (transcriptions of man-man dialogues or of Wizard of Oz experiments<sup>1</sup> etc.) but transcription practices are so different that it is impossible to define unified exploration tools which would allow the research community to fully exploit them.

From considering linguistic resources as generally available data we are led to consider the tools associated with them. These tools will change according to the category of user: transparent, data integrated (on-line tools) or widely distributed and adaptable tools (e.g. programming libraries/API<sup>2</sup> ).

Clearly there is a lot to consider and all difficulties cannot be solved at once. But the CNRS<sup>3</sup> and the Aupelf•Uref<sup>4</sup> have been looking to improve the re-usability of data within the French-speaking community. This is being spearheaded by a joint venture bringing together 5 academic teams<sup>5</sup> to address a maximum of French-speaking laboratories and other sites.

---

<sup>1</sup> These experiments simulate a man-machine dialogue system which aims to observe the "spontaneous" behaviour users could have in front of such systems.

<sup>2</sup> Application Programming Interface

<sup>3</sup> Centre National de la Recherche Scientifique

<sup>4</sup> Association des Universités Partiellement ou Entièrement de Langue Française

<sup>5</sup> Loria (Nancy), INaLF (Nancy, Paris), LPL (Aix), LIMSI (Orsay) and CLIPS (Grenoble)

This paper presents a synthesis of the deliberations which have led to the implementation of the first experimental server of linguistic resources.

## **2.General Objectives.**

SILFIDE (Serveur Interactif pour la Langue Française, son Identité, sa Diffusion et son Etude) is a platform designed for the dissemination of standardized data and tools relating to the French language. The overall objective is to introduce a network of data processing servers and support systems. The aim of SILFIDE is not necessarily just to integrate all the contents of available resources (corpora, glossaries, tools), but also to inform researchers of the existence of such contents, get a relatively precise idea of them, and be aware of the possible means of access.

In the case of general purpose resources which do not give rise to legal problems, SILFIDE will be able to provide the automatic transfer of the corresponding data. Past and present initiatives (e.g. Consortium for Lexical Research, (<http://crl.nmsu.edu>) or the European Linguistic Resource Association, (<http://www.icp.grenet.fr/ELRA/home/html>) have set out to achieve similar objectives, but have only acted as a clearing house and repository, gathering linguistic resources whatever their format and distributing them as they are either through an ftp site or on CD ROMs. SILFIDE goes further by putting emphasis both on high quality encoded data and on specific access tools for on-line work.

Furthermore, SILFIDE is a help tool provided to all the laboratories in the French-speaking community and to those who are interested in the study or the automatic treatment of the French language. For this reason, French must remain the principal language of the project. Most of the data available on SILFIDE will be in French or associated to equivalent data in French (e.g. in the case of a parallel corpus). French will also be the meta-language for the management of the resources both for their documentation and at the level of the access interface to the corpus. However, a description of the server in other languages (e.g. English or German) would be useful.

Given the importance of sharing expertise in the field of standardized data delivery, the underlying technology must be kept generic enough (e.g. through the use of Unicode) to make sure that it can be duplicated at any site.

### **Main Functions**

At the beginning, SILFIDE should be able to answer the following potential questions :

- What are the available data ?
- Where are they available and under which format ?
- What are the conditions of access ?
- How (and by whom) have these data been compiled ?
- What is the validation degree of the resources ?
- What are the tools available to manipulate these resources ?

#### **Other functions**

Besides the function of access to linguistic resources, directly accessible 'on-line' tools may also help users who do not have access to an elaborate computer environment. Concordances can then be developed for a set of selected texts together with elementary lexical statistics (frequencies, reduced deviations, etc.).

SILFIDE should also aim to compile and/or document the tools available in the field of textual resources manipulation. It can be a matter of encoding data but also library functions dedicated to normalized data. These different additional functions will have to be progressively integrated to the successive versions of the SILFIDE server.

### **3. Encoding data into the TEI scheme: why? and how?**

Considering the different objectives of SILFIDE, the project relies on an underlying framework for the representation of structured documents in an electronic format.

It was logical to follow the Text Encoding Initiative (cf. TEI -P3) rather than devise another scheme. However, it has been necessary to simplify and even misuse the actual guidelines provided by the TEI.

In such a large and multifarious project, the TEI will be considered in different ways by the different interest groups involved, which in turn will depend on :

- a) the different data types that are to be represented, and
- b) the different possible usages.

### 3.1 The librarian vs. the linguist viewpoint on data

From the outset, SILFIDE had to accommodate two opposing views of data to be distributed. On the one hand it had to provide the user with a concise and accurate description of the available linguistic resources which could be queried easily and rapidly. On the other hand, it had to provide specific categories of users with on-line access to the actual content of any resource for specific research purposes. The conflict between efficiency and exhaustiveness was solved by clearly assigning two different functions to the TEI header (*teiHeader* element) on one side and the content proper (*text* element) on the other. Accordingly, a user scenario was devised which relies mainly on two phases, one during which the user selects the resources he wants to work on, and puts them into a “shopping basket”, and one where the actual work on the resources takes place using specific tools which actually use the structural content of the data.

Thus the TEI header functions like a user-friendly database of precisely identified fields (title, author, bibliographic source, etc.) and no specialist TEI knowledge is required. To this end, the (virtual) set of headers associated with the whole data fund was compiled into a database accessible through a set of indexes directly queryable by the user.

However, putting this into practice may prove difficult given the degree of (albeit useful) flexibility of the TEI vis-à-vis the precise structure of the header. Designing a single indexing scheme upon the header was all the more difficult due to quite a large variety of document types or genres; (from “standard” narrative texts and plays to transcriptions of oral documents, dictionaries and lexica) all of which require specific variations of the TEI header. For instance, the transcription of an oral dialogue implies an extensive and detailed use of both the *sourceDesc* element (via the *recordingStmt* element) and the *profileDesc* element (particularly via the *particDesc* element), which are used quite differently in the case of, say a novel.

With this in mind, we chose to adopt the following editorial policy regarding the header :

- to impose a set of fields are mandatory and are to be shared by all the resources in SILFIDE;
- to allow a great flexibility for the other fields to account for possible variations between genres.

In this way, attention was primarily focused on the description of two of the four sections which make up the TEI header : the *fileDesc* element (dedicated to the description of the electronic file and its possible source) and the *profileDesc* element (describing the informational characteristics of the resource). In *fileDesc*, the following information is required : title, author and a basic responsibility statement (in particular to trace who made the resource available to SILFIDE), a minimal description of the extend (in free format, in general number of words, dialogues or newspaper articles), a general publication statement for the distribution of the resource, and a description of the source which depends on the nature of the resource (e.g. a bibliographical description of the textual reference using the *biblStruc* element). In the same way, our use of *profileDesc* is centred on fields which are to be automatically indexed within the server, such as *langUsage* (for the description of the languages represented in the content), *textDesc* (in particular to indicate whether the resource corresponds to spoken or written language), *particDesc* (to describe the characters in a play, dialogue, etc) and *textClass* (to describe a set of generic keywords associated with the resource).

### 3.2 Accessing content

The difficulties are not limited to providing a sound and generic description of the header for all resources. There have also been difficulties in devising a clear editorial policy concerning the way the actual content of data is encoded. As observed by several encoding projects which have used the TEI (e.g. the Women Writers Project at Brown), there is always a compromise to be reached between a) the precision of the encoding which should be as refined as possible and b) the level of genericity of the corresponding document, i.e. its compatibility with different possible usages. The key factor associated with this compromise is that if one wants to keep a homogenous encoding scheme within a database, each step towards more refined encoding may prove not only costly and highly time-consuming, but also difficult to control and maintain.

As a result, the following general principles were adopted to encode resources.

- **Identifying the structure** - each resource is encoded so that the basic structure of its content can be retrieved. For textual resources, this means the whole

hierarchy of divisions<sup>6</sup> , and a basic representation into paragraphs or a similar level of description (<u> for the transcription of speech utterances, <sp> for play parts, <lg> for embedded poetry stanzas, in which case <l>s were used to keep line information;

- **Fidelity to the electronic source** - since our data are already in an electronic format when they reach us<sup>7</sup> , any elementary feature which is present in the source file is encoded. Thus any typographical markers are converted into meaningful elements for textual documents, or keep specific transcription indications for spoken data (e.g. <pause>, <unclear>, <note> etc.);

- **Improving data for specific use** - depending on the tools the user requires in order to work on the data, encoding must be broken down to the level of the specific elements on which it relies. For example, take the specific case of the parallel alignment of a text and its translation which we systematically conduct up to the sentence level (Romary et al., 1995). In this case, for each text which could be associated with a translation in any other language, all versions are semi-automatically sliced into either <s>s (in particular, when using the TEI Lite DTD) or <segs> (in the case where there were numerous interferences with existing <q>s).

Conformity to TEI P3 (e.g. by selecting the proper options in the DTD), has been a constant aim of the project, rather than modifying the system to one which might not be shared by other similar projects or potential users.

#### 4. *Modus operandi*

There is no need to detail here the technical platform from which the current version of the SILFIDE server is derived. We can simply point out that all the developments are based on the Internet network and its protocols, so that ultimately it would be possible to have direct access to the server from any standard web browser. However, the SILFIDE service, unlike some private initiatives such as the ABU server (Association des Bibliophiles Universels, <http://cedric.cnam.fr/ABU/>), is not intended for the general public but for a community of researchers who wish to work

---

<sup>6</sup> We might mention here that using <div>s for encoding divisions has proven to be far more flexible than using numbered divs (div1, div2, etc.).

<sup>7</sup> As opposed to many projects using the TEI around the world, SILFIDE is not supposed to create new electronic resources *per se*, but rather put together ones which already exist in the academic community.

on the French language. To avoid making the procedure too heavy, we have set up a registering system which identifies the different users of the server as suppliers or mere users/readers.

Consequently, all the functions of the server which require direct access to the resources themselves are not available without prior authorization.

In outline, the SILFIDE server takes the form of a navigation which gives access to the following functions :

- General information about the server itself
- Access to resources through navigation (title or author for instance, in the case of literary texts) or through a more complete search;
- A set of service functions, in particular standard tools which are available on-line or on free access;
- The possibility to register as a user (or supplier)
- an interactive mode with the server itself in order to provide additional information, comments, etc.

The SILFIDE server, which in its experimental version currently contains an initial corpus of texts and dialogue transcriptions (about 5 million words for 30 megabytes of data) is accessible at <http://www.loria.fr/projets/SILFIDE>.

## 5. Outlook

The SILFIDE project will only prove its usefulness when it becomes a "natural" component of research requiring linguistic data, i.e. a site which a user will spontaneously and systematically think of using to search for the data necessary to his work. This 'usefulness' will also be proven when he feels like a potential supplier. At first, and in accordance with the initial objective of the project, SILFIDE must accompany the structuring actions of our community, such as the Concerted Research Actions of the Aupelf•Uref. Beyond this point, it is important for the project to be enriched by related developments, both in terms of its contents (data, tools) and in the scope of research projects which would rest on this structure. Finally a medium term perspective is the transmission of the SILFIDE model to other sites in Europe or wherever a similar server is required for languages other than French, or within the context of specific projects (e.g. structuring actions with Eastern Europe).

An enrichment of the structure is conceivable, because the compatibility of the different fields should ultimately allow for the interconnection of such servers, and also because each site could develop additional access tools available to all. An even more ambitious prospect is now for us to see SILFIDE as a general purpose delivery tool for structured documents of any kind. We are currently studying the feasibility of delivering gene and protein sequences on the basis of existing XML based encoding schemas such as BSML (Bioinformatic Sequence Markup Language, <http://www.topogen.com/bsml>). In fact both textual and biological documents can be seen as semi-structured data which can be queried and accessed in very similar ways (see Buneman et alii, 1996).

Finally putting the TEI into practice clearly shows that in addition to aiming at being a “standard”, the TEI is not only an occasion to share common practices, but also an opportunity to share a kind of philosophy in the encoding of textual documents. Above all, TEI will actually prove valuable when we will really be able to exchange both data and tools (such as SILFIDE) without having to revise either of them. This is something which is not yet attainable, but can be achieved by even more collaborative work between the sites which are concerned by digital resources.

## 6. References

- Aarts Jan, Peter de Haan et Nelleke Oostdijk (Eds), *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, Amsterdam, 1993.
- Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC) (1994), *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*, Editions C. M. Sperberg-McQueen and Lou Burnard, 2 volumes, Chicago, Oxford: Text Encoding Initiative.
- Buneman P, Semistructured Data, Tutorial PODS '97 (see <http://www.cis.upenn.edu/~db/tutorials.html>).
- Dunlop D. (1995) Practical Considerations in the Use of TEI Headers in a Large Corpus, *Text Encoding Initiative: Background and Context*, Kluwer Academic Publishers, Dordrecht, p. 85-98.

- Heid U. and Oliver C. (1996) *An Investigation into the Use of AFS for distribution and networking of linguistic resources and tools*, Technical Report Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.
- Ide N. et Véronis J. (1995) *MULTEXT/EAGLES-Corpus Encoding Standard*, document Version 0.1. CNRS, Aix-en-Provence.
- IJCL-96, *International Journal of Corpus Linguistics*, V. 1 N.1, John Benjamins, 1996.
- Krol E. (1992), *The Whole Internet: user's guide and catalog*, O'Reilly & associates, Sebastopol, collection Nutshell Handbook.
- Lapeyre D.A. & Usdin T. (1996) *TEI and the American Memory Project at the Library of Congress*, Workshop : The Text Encoding Initiative Guidelines and their Application to Building Digital Libraries (20-23 March 96)
- Péry-Woodley Marie-Paule, « Quels corpus pour quels traitements automatiques ? », in TAL-95.
- Pino M. (1996) *Encoding two large Spanish corpora with the TEI scheme: design and technical aspects of textual markup*, Workshop : The Text Encoding Initiative Guidelines and their Application to Building Digital Libraries (20-23 Mars 96)
- Romary Laurent, Nathalie Mehl and David Woolls 1995, *The Lingua Parallel Concordancing Project: Managing Multilingual Texts for Educational Purpose*, *Text Technology*, 5, 3, pp. 206-220.
- TAL-95, *Traitement probabilistes et corpus*, *Traitement Automatique des Langues* journal, Volume 36, Number 1-2, 1995.
- TEI-P3, Association for Computers and the Humanities (ACH), Association for Computational Linguistics (ACL), and Association for Literary and Linguistic Computing (ALLC) (1994), *Guidelines for Electronic Text Encoding and Interchange* (TEI P3), Editions C. M. Sperberg-McQueen and Lou Burnard, 2 volumes, Chicago, Oxford: Text Encoding Initiative.