Le projet Silfide : vers un accès ouvert aux ressources linguistiques francophones

Laurent Romary, Jean-Marie Pierrel

Le projet Silfide : vers un accès ouvert aux ressources linguistiques francophones[1]

Laurent Romary
Jean-Marie Pierrel
CRIN-CNRS & INRIA Lorraine
Bâtiment Loria, B.P. 239, F-54506 Vandœuvre Lès Nancy
{romary,jmp}@loria.fr

1. Introduction

Ces dernières années ont vu naître un regain d'intérêt pour des études reposant sur des ressources linguistiques informatisées, tant de point de vue des sciences humaines pour des études linguistiques, littéraires ou historiques que de celui de l'informatique. Différentes publications récentes attestent de la vivacité du mouvement (par exemple Aarts et alii 92, TAL-95, IJCL-96), mais surtout de la pluralité des méthodes et des objectifs employés par les chercheurs du domaine. De fait, ce renouveau des méthodes pose un ensemble de questions de fond quant au statut et à la maintenance des données ainsi manipulées. En effet, il ne semble plus possible de répéter à l'infini le cycle de travail sur des données que l'on a presque toujours observé dans nos communautés : dans le cadre d'un projet de recherche particulier, on définit les données qui seraient nécessaires, puis on effectue un recueil de celles-ci, avant de construire rapidement deux trois outils *ad hoc* qui permettront d'extraire les informations pertinentes pour l'étude en cours. Enfin, quand tout est fini et que les résultats de recherche sont publiés, les données sont laissées à elle-même sous une forme plus ou moins identifiée, et surtout subordonnée à l'existence d'une mémoire collective des participants au projet de recherche initial. Dans la plupart des cas, ces données deviennent complètement inutilisables par tout autre projet du même type, soit parce qu'il n'existe plus d'outil informatique compatible, soit parce que les formats de représentation n'ont pas été documentés, soit encore parce qu'il serait trop coûteux de transformer ces donnés de manière à les rendre compatibles avec les nouveaux outils définis pour la recherche en cours. De fait, ce dernier point explique aussi en partie l'impossibilité d'envisager jusqu'à maintenant une utilisation souple et modulaire de données issues de grands fonds textuels, dont les formats spécifiques n'ont jamais été associés à la distribution d'outils disponibles largement au sein des communautés académiques.

---

[1] Cet article doit paraître dans un numéro spécial de la Revue Française de Linguistique Appliquée, 1997.

Très clairement, le problème qui se pose à nous ici est celui de la réutilisabilité. Ce problème ne peut en soit être posé de façon général et il faut essayer de définir quelques axes de réflexion pouvant conduire à des réponses réalistes pour notre communauté. Tout d'abord, il faut distinguer les différentes ressources linguistiques que l'on souhaite représenter. Le cas des données textuelles semble de prime abord le plus simple, de part le faible degré de structuration que celles-ci impliquent. On observera cependant que même dans les modes de représentation les plus simplifiés (texte non balisé), il est nécessaire d'adjoindre un minimum de documentation quant à l'origine et au contenu des textes correspondants. Par ailleurs, suivant en cela l'avis de M.-P. Pery-Woodley (95), il nous semble essentiel de recueillir les données textuelles au niveau de textes complets et identifiables de façon à maîtriser parfaitement tous les paramètres (genre, structure) susceptibles d'être utilisés pour des études ultérieures. Le texte doit être vu comme une entité comptable et non pas massive…

Les autres ressources linguistiques sont en général, par essence, beaucoup plus structurées et nécessitent de ce fait d'autant plus d'attention si l'on souhaite les rendre disponible à une large communauté. Parmi celles-ci, on pourra mentionner le cas des ressources lexicales qui prendront, suivant l'usage que l'on souhaite en faire, la forme d'un dictionnaire informatisé (pour une consultation humaine) ou d'une véritable base de données lexicales (pour une utilisation automatisée). Dans ce dernier cas il est indispensable de normaliser pleinement la structure utilisée, de sorte que ces ressources puissent être intégrées dans différentes plates-formes de développement. De la même façon, il existe actuellement de nombreux corpus de dialogues (transcriptions de dialogues homme-homme, résultats d'expériences de type Magicien d'Oz[2] etc.), mais sous une telle disparité de formes qu'il est impossible de définir des outils unifiés d'exploration qui rendent ces corpus véritablement exploitables. On le voit, partant des ressources linguistiques prises au sens des données disponibles, on aboutit vite aux outils que l'on souhaite associer à celles-ci. De fait, mettre à disposition des données doit, au risque de ne rester qu'un voeu pieux, s'accompagner d'une réflexion approfondie sur les environnements de travail qui permettent de les manipuler. Ainsi, suivant la catégorie d'utilisateurs, on souhaitera des outils

[2] Il s'agit d'expériences de simulation de systèmes de dialogue homme-machine dont l'objectif est d'observer le comportemlent « spontané » que pourraient avoir des utilisateurs devant de tels systèmes.

transparents, quasi-intégrés au données (outils en ligne), ou des environnements (librairies informatiques par exemple) largement distribués et modulables.

Comme on peut le constater, la réflexion à mener est d'importance, et il est clair que toutes les difficultés ne peuvent être résolues d'un seul coup. C'est cependant dans l'espoir d'une amélioration de la situation francophone en la matière que le CNRS et l'Aupelf•Uref ont lancé une initiative conjointe qui, réunissant dans un premier temps 5 équipes universitaires[3], doit à terme s'adresser à un maximum de sites ou de laboratoires francophones. Dans cet article, nous présentons une synthèse des réflexions qui ont mené à la réalisation d'un premier serveur expérimental.

2. Objectifs généraux

SILFIDE (Serveur Interactif pour la Langue Française, son Identité, sa Diffusion et son Etude) est un outil de mise en commun, conviviale et raisonnée, des connaissances sur différents aspects de la langue française. Il consiste en un réseau de serveurs informatiques et d'actions en alimentant les fonctions.

SILFIDE n'a pas en soi vocation d'intégrer l'ensemble des contenus (corpus, lexiques, outils) des ressources disponibles au sein de la communauté universitaire, mais bien de permettre à tout chercheur d'avoir connaissance de l'existence de ces contenus, de s'en faire une idée relativement précise et de connaître les modalités d'accès à ceux-ci. Dans le cas de ressources d'usage très général, ou ne posant pas de problème spécifique d'accès, SILFIDE pourra directement proposer le transfert des données électroniques correspondantes.

Un serveur francophone. SILFIDE est un service rendu à l'ensemble des laboratoires de la communauté francophone ou s'intéressant à l'étude ou au traitement automatique de la langue française. A ce titre le français se doit d'être une langue centrale de notre projet. D'une part, l'essentiel des données disponibles sur le serveur SILFIDE sera en français ou associé à des données équivalentes en langue française (par exemple dans le cas de corpus parallèles). D'autre part, le français sera la métalangue associée à la gestion des ressources, que ce soit au niveau de la documentation de celle-ci, ou de l'interface d'accès au corpus. Il serait malgré tout utile qu'une

[3] Le CRIN (Nancy), l'INaLF (Nancy/Paris), le LPL (Aix), le LIMSI (Orsay) et le CLIPS (Grenoble)

description du serveur soit disponible dans d'autres langues de travail (anglais ou allemand par exemple).

Fonctions principales. SILFIDE doit dans un premier temps pouvoir répondre aux interrogations suivantes de la part d'un utilisateur :

- Quelles sont les données disponibles ?
- Où celles-ci sont-elles accessibles et sous quel format ?
- Quelles sont les conditions d'accès ?
- Dans quelles conditions ces ressources ont-elles été compilées (et par qui) ?
- Quel est le degré de validation des ressources ?
- Quels sont les outils disponibles pour manipuler ces ressources ?

Autres fonctionnalités - Au-delà des fonctionnalités d'accès à des ressources linguistiques, il peut être important de proposer à certains utilisateurs ne disposant pas d'environnement informatique élaboré d'autres outils directement accessibles « en ligne ». On peut ainsi penser à des concordances sur un ensemble de textes qui aura été sélectionné, accompagnées de statistiques lexicales élémentaires (fréquences, écart réduit etc.).

Par ailleurs, SILFIDE doit pouvoir rendre une fonction de service en recensant (et éventuellement en documentant en français…) les outils disponibles dans le domaine de la manipulation de ressources textuelles. Il peut s'agir tout aussi bien du codage des données, mais aussi des librairies de fonctions dédiées aux données normalisées. Ces différentes fonctionnalités supplémentaires devront progressivement être intégrées aux versions successives du serveur SILFIDE.

3. Nature des données

3.1. Normalisation des données

L'expérience acquise au sein des différents serveurs de ressources linguistiques, en Europe comme aux USA, montre qu'il est nécessaire de représenter les données rendues disponibles sous une forme qui soit relativement normalisée et ce pour tout un ensemble de raisons :

a) En premier lieu il est nécessaire de pouvoir identifier de manière précise l'origine et la nature des données proposées. Ceci concerne tout d'abord la description bibliographique précise lorsqu'il existe une source sous forme papier, ou le contexte d'expérimentation dans le cas de données issues de retranscriptions de corpus oral. Il faut aussi pouvoir déterminer le degré de validation des données (relectures successives pour des données scannérisées par exemple, ou conditions de la transcription).

b) Il est par ailleurs important qu'un utilisateur puisse avoir une idée *a priori* du mode de codage qui lui est proposé (codage des caractères, codage de la structure) sous peine de devoir lui-même identifier ces paramètres, ressource par ressource.

c) Enfin, et ceci est intimement lié au point précédent, un certain niveau de normalisation permet d'envisager à terme de mettre à disposition des utilisateurs des outils standardisés de manipulation des ressources directement depuis le serveur. D'ores et déjà, le succès de la TEI[4], sur une base SGML, a permis de voir fleurir des bibliothèques d'outils ou des environnements logiciels supportant tout ou partie des directives correspondantes (éditeurs SGML bien sûr, mais aussi des API (Application Programming Interface/Interface Applicative de Programmation) telles que celles développés à Nancy autour de la Dilib, à Edimbourg, ou des langages de requêtes tels que SGMLQL à Aix).

Une remarque importante à faire ici est qu'il est nécessaire, lorsque l'on aborde le problème de la normalisation des données, de faire une distinction entre usage privé et usage publique d'une ressource linguistique. Ces deux usages vont en effet se différencier par une pratique relative au codage. D'un coté, l'usage privé aura tendance à négliger la documentation des ressources (leur origine en est en général bien connue localement) pour au contraire favoriser une élaboration des marques de structuration et d'annotation de celles-ci (les ressources étant destinées à des traitements bien particuliers[5]).

---

[4] SGML : *Standard Generalized Mark-up Language* est un formalisme de description de documents structurés, bien adapté en particulier aux documents textuels. Ce standard repose sur la définition d'une syntaxe de document ou DTD (*Document Type Definition*). La TEI est un ensemble de directives issues d'un travail de réflexion de plus de huit ans au sein de la communauté informatique et sciences humaines pour la représentation de nombreux types de documents (prose, poésie, théatre, dictionnaires, lexiques etc.). Ces directives peuvent être vues comme une DTD particulière.

[5] Les équipes travaillant sur l'alignement multilingue vont par exemple systématiquement reconstituer la

Une autre différence, qui est d'ailleurs explicitée par la TEI, concerne, d'une part, l'usage d'un jeu de caractère plus réduit dans le cadre d'une ressource à usage publique (en complétant les caractères manquant à l'aide d'entités SGML) et, d'autre part, l'application de règles plus strictes quant au codage des éléments SGML (par exemple en insérant systématiquement les balises finales dont l'usage serait optionnel dans la DTD correspondante).

Si nous nous plaçons maintenant sur le plan de la documentation des ressources, on peut remarquer que celle-ci (que la ressource soit de nature textuelle, orale (transcription) ou lexicale) doit comporter au minimum les éléments suivants :

- description de la ressource électronique elle-même ;
- description bibliographique précise ;
- description des pratiques de codage (codage de référence, par exemple le niveau de segmentation maximal, ainsi que la liste des balisesutilisées y seront indiquées) ;
- l'histoire du document, en particulier les différentes corrections que celui-ci a subies, de façon en particulier à pouvoir évaluer la qualité de la ressource (notion de validation).

Enfin, il faut que d'une manière ou d'une autre cette documentation accompagne systématiquement toute ressource linguistique de manière d'une part à ne pas aboutir à des ressources qui ne soient pas répertoriée et d'autre part à ce qu'aucune contradiction n'apparaisse entre une ressource donnée et sa documentation.

Ces différents éléments ont été pris en compte dans les directives de la TEI dans le cadre de la partie "en-tête" (TEIheader) du document SGML codant la ressource correspondante. Une structure d'en-tête simplifiée et adaptée aux ressources propres à l'ingénierie linguistique a été proposée dans le cadre de la CES[6]. Le rapport de spécification du projet Silfide détaille les formats spécifiques de l'en-tête TEI, associés à l'usage que feront respectivement les fournisseurs de données et les utilisateurs du serveur. Il est clair que toute ressource se conformant à ces consignes minimales pourra être documentée de façon bien plus précise, de sorte à la rendre

---

structure en phrases de leurs ressources, alors qu'une telle opération (nécessairement imparfaite) peut être préjudiciable à quelqu'un qui souhaite faire une étude sur la ponctuation.

[6] Corpus Encoding Standards - Il s'agit d'un ensemble de directives issues notamment des travaux menés au sein du projet européen Eagles.

encore plus exploitable par une majorité d'utilisateurs potentiels (cf. la notion d'usage public d'une ressource ou document).

Ainsi donc, le codage des textes se basera sur les recommandations établies dans la CES, qui est une application de SGML et conforme aux directives de la TEI (*TEI Guidelines*). Tout texte fourni devra être :

- accompagné d'un document présentant les caractéristiques du texte (informations sur versions électroniques, texte source...), c'est le teiHeader (en-tête TEI). Cet en-tête sera créé à partir des informations contenues dans les fiches d'identification des textes complétées par le fournisseur.
- au moins codé sous une forme minimale pour pouvoir l'insérer à l'ensemble des ressources disponibles sur le serveur (les CES décrivent d'ailleurs différents niveaux de codage).

Un premier bilan

Lorsque l'on parle de norme(s), il est clair qu'il ne faut pas espérer disposer dans l'immédiat d'un ensemble exhaustif, et surtout cohérent, de directives permettant de représenter tout type de ressource linguistique, du glossaire médiéval manuscrit jusqu'à la bande dessinée contemporaine. Il faut reconnaître cependant le travail énorme effectué au sein des différents groupes de travail de la TEI en particulier pour couvrir un champ très large des besoins en sciences humaines comme en informatique (avec le complément de réflexion introduit par les CES). Dans le cadre de la première phase du projet Silfide, nous avons mené différentes expérimentations qui valident en bonne partie les directives existantes, et tout particulièrement, le choix de s'appuyer sur SGML comme cadre général. Ainsi nous avons été amené à travailler sur :

- des textes littéraires, avec en particulier une expérience de rétroconversion de certain textes issus de la base Frantext de l'INaLF ;
- des transcriptions de dialogues homme-homme (en situation réelle ou résultant d'expériences dites du Magicien d'Oz). Dans ce cadre, des propositions de directives plus fines ont été faites à la CES ;
- des lexiques. Nous pouvons citer ici le travail effectué sur le dictionnaire de Basnage (18ème siècle) qui peut très facilement être intégré dans la forme la plus standard de la TEI.

3.2. Texte intégral vs vitrine

Comme nous l'avons signalé, les données présentées par le serveur SILFIDE peuvent être aussi bien des données directement disponibles dans le serveur, que des pointeurs sur des ressources accessibles sur d'autres sites. Quoi qu'il en soit, il est indispensable que ces deux types de données puissent être représentées de façon similaire, de manière à donner une vision uniforme des ressources disponibles à un utilisateur. Cette uniformisation est rendue effective d'une part par l'usage d'un format unique de documentation de ces données, et d'autre part par la mise à disposition systématique d'une « vitrine », où des échantillons issus de la ressource sont proposés à l'utilisateur. La systématisation d'une telle vitrine prend en compte différentes contraintes :

- La limitation d'accès à certaines ressources, du fait des contraintes de droit. La vitrine doit alors respecter les limitations juridiques applicables à la ressource (par exemple le nombre de lignes de l'édition d'origine d'un livre mis sous forme électronique) ;
- L'existence de nombreuses ressources qui ne se présentent pas sous une forme normalisée ou même qui ne sont accessibles qu'à l'aide d'un logiciel particulier (par exemple hypercard). Il faudra alors prévoir une petite opération de codage pour les échantillons correspondants.

La vitrine d'une ressource textuelle pourra contenir différents types d'informations :

- une présentation du contenu sur la base des informations présentes dans l'en-tête (taille de la ressource, circonstance de sa création, locuteurs etc.),
- une vue globale de sa structure (par exemple la structure en chapitres d'un texte, quand cette information a été codée),
- des échantillons de contenu (ensemble de paragraphes désignés par le fournisseur comme pouvant être rendu publics).

4. Mode de fonctionnement

Il n'est pas nécessaire de détailler ici la plate-forme technique sur laquelle repose la version actuelle du serveur Silfide. Nous pouvons simplement signaler que l'ensemble des développements repose sur le réseau Internet et ses protocoles de sorte qu'à terme, il devrait être possible d'accéder au serveur à partir de n'importe quel lecteur de web standard. Cependant, le serveur Silfide, contrairement à certaines initiatives privées telle que le serveur de l'ABU

(Association des Bibliophiles Universels), n'est pas destiné au grand public en général, mais bien à une communauté de chercheurs souhaitant travailler sur la langue française. Sans rendre la procédure particulièrement lourde, nous avons donc mis en place un système d'enregistrement, qui permet d'identifier les différents utilisateurs du serveur, qu'ils soient fournisseurs, ou simple « lecteur ». Dès lors, toute les fonctionnalités du serveur qui requièrent un accès aux ressources proprement dites ne sont disponibles qu'après une autorisation explicite.

Dans ses grandes lignes, le serveur Silfide se présente sous la forme d'une navigation donnant accès au fonctionnalités suivantes :

- des informations générales sur le serveur lui-même ;
- un accès aux ressources par navigation (par exemple titre, auteur etc. dans le cas de textes littéraires) ou par une recherche plus complète ;
- un ensemble de fonctions de service, notamment des outils standards disponibles en ligne ou en libre accès ;
- la possibilité de s'inscrire comme utilisateur (et éventuellement fournisseur) ;
- un mode d'interaction avec le serveur lui-même, pour apporter de l'information supplémentaire, des commentaires etc.

Le serveur Silfide, qui dans sa version expérimentale contient actuellement un corpus initial de textes et de transcription de dialogues (environ 5 millions de mots, pour 30 Méga-Octets de données), est accessible depuis l'adresse suivante : http://www.loria.fr/Projet/Silfide

5. Perspectives

Le projet Silfide n'aura prouvé son utilité qu'à partir du moment où il deviendra une composante « naturelle » de toute recherche reposant sur des données linguistiques, c'est à dire un lieu où spontanément un utilisateur songera systématiquement à rechercher les données nécessaires à son travail sur le serveur, et que chacun se sentira l'âme d'un fournisseur potentiel. Dans un premier temps, et conformément aux objectifs initiaux du projet, Silfide doit accompagner les actions structurantes de notre communauté, telles que les Actions de Recherche Concertées de l'Aupelf•Uref. Au delà, il est important que le projet s'enrichisse de développement connexes, tant du point de vue de son contenu (données et outils) que dans le cadre de projets de recherche qui s'appuieraient sur sa structure. Enfin, une perspective à moyen terme semble être de développer le modèle Silfide pour le transmettre à d'autres sites en Europe

ou ailleurs souhaitant développer un serveur similaire pour d'autres langues que le français, ou dans le cadre de projets spécifiques (on peut penser à des actions structurantes avec l'Europe de l'Est par exemple). De la sorte, il est imaginable d'aboutir à un enrichissement de la structure, d'une part parce que la compatibilité des domaines doit permettre à terme l'interconnexion de tels serveurs, et d'autres part parce que chaque site peut développer des outils d'accès supplémentaires utilisables par tous.

6. Références

Anuff E.(1996) The Java Sourcebook, J. Wiley and sons, New-York, Chichester, Brisbane.

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.

Béthery A. (1993) *Abrégé de la classification décimale de Dewey*, Editions du cercle de la librairie, Collection Bibliothèques.

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 ressources 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 Landbook.

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 Mars 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 Librairies (20-23 Mars 96)

TAL-95, *Traitement probabilistes et corpus*, revue t.a.l., Volume 36, Numéro 1-2, 1995.