Une approche reposante (RESTful) des aspects opérationnels de la rétroconversion du FEW

Cyril Briquet (1), Pascale Renders (1) (2)
(1) ATILF (CNRS & Nancy-Université), (2) Université de Liège
cyril.briquet@acm.org, pascale.renders@ulg.ac.be

Résumé

Les articles du *Französisches Etymologisches Wörterbuch* (FEW) seront bientôt enrichis d'un balisage (XML) sémantique par une séquence d'algorithmes. Nous proposons un modèle éditorial distribué (“à la Wikipedia”) des corrections manuelles à apporter aux articles qui ne peuvent être automatiquement balisés. En outre, de manière à rendre totalement transparent pour les contributeurs le maintien de la consistance et de la synchronisation des trois quarts de million de documents considérés, nous proposons une plate-forme web de rétroconversion basée sur une architecture logicielle reposante (*RESTful*, cf. article Wikipedia sur REST).

1 Introduction

Le problème de *rétroconversion* consiste à identifier automatiquement les types d'information importants d'un dictionnaire papier puis à insérer un balisage (qualifié de sémantique) marquant leur position dans le texte. Ce problème, en particulier pour le *Französisches Etymologisches Wörterbuch* (FEW), dictionnaire étymologique et historique des langues de la Galloromania, est complexe en raison de la nature souvent implicite des informations (unités lexicales, étymons, datations, rédacteurs,...) et de leur manque de consistance syntaxique, ainsi qu'en raison du grand nombre (~20000) d'articles. Un traitement algorithmique de la version plein texte des articles numérisés garantit une consistance certaine et promet un gain de temps appréciable. Toutefois, il n'est jamais complètement fiable en raison des nombreuses irrégularités de ce type de source de données.

Tout en automatisant ce qui est automatisable, un recours à l'expertise humaine permet de régulariser - en vue de leur réappliquer un traitement automatique - les données présentant des irrégularités non prévues. Vu le grand nombre d'articles à rétroconvertir, il est prudent et important de prévoir une organisation efficace du processus de corrections manuelles ainsi que de ses interactions avec l'exécution automatique des algorithmes de rétroconversion.

De ce point de vue, plusieurs problèmes doivent être résolus. Premièrement, un système de gestion des versions d'un article en cours de rétroconversion doit être mis en place, vu le grand nombre d'articles concernés. Deuxièmement, un système d'échange de données entre plusieurs experts humains coopérant au processus de corrections manuelles doit être instauré, sans quoi les échanges non structurés de fichiers, e.g. via e-mail ou clé USB, deviendront rapidement source d'erreurs. Troisièmement, un système de conservation à long terme des données et des corrections doit être mis à disposition des équipes d'experts humains qui se succéderont et coopéreront au fil du temps.

La rétroconversion du FEW est à la fois un problème classique (cf. TLFi et OED) et une entreprise ambitieuse. L'originalité de notre approche de ses aspects opérationnels est de recourir de manière appropriée à l'automatisation ainsi qu'à l'expertise humaine, afin que les avantages qui en découlent soient plus nombreux que ceux qu'on pourrait espérer si on y avait recours séparément. Concrètement, nous proposons, d'une part, de traiter algorithmiquement les données des articles numérisés et, d'autre part, d'adopter un modèle éditorial coopératif à large échelle (“à la Wikipedia”) en organisant via une plate-forme web à accès contrôlé les corrections manuelles réalisées par les experts du FEW.

L'approche que nous proposons repose sur ces éléments: *(i)* modèle éditorial spatialement et temporellement distribué (cf. Wikipedia) des articles algorithmiquement rétroconvertis, *(ii)* plate-forme basée sur le web garantissant la consistance et la synchronisation des données et *(iii)* style architectural (i.e. REST) fluidifiant l'accès aux articles ainsi que l'activation des opérations que l'on peut leur appliquer; le tout permettant sans effort supplémentaire l'agrégation d'une multitude de petites contributions éditoriales. Un prototype logiciel (pas encore public) réalise déjà une grande part de notre proposition.

Le reste de l'article est structuré comme suit: la section 2 résume le processus de rétroconversion du FEW, la section 3 introduit la plate-forme de rétroconversion distribuée que nous proposons et la section 4 discute des perspectives d'évolution.

2 Processus de rétroconversion

2.1 Objectif de la rétroconversion

L'objectif de la rétroconversion du *Französisches Etymologisches Wörterbuch* (FEW) est d'en obtenir une version informatique comportant un balisage sémantique identifiant les types d'information importants: unités lexicales, étymons, datations, affixes, étiquettes géolinguistiques, rédacteurs,... Les données d'entrée sont constituées de documents XML. Chaque article du dictionnaire est représenté par un document XML au format FEW Font-style Markup Language (FFML) comportant un balisage de formatage: gras, italique, exposant, petites capitales, paragraphes, sauts de colonne et de ligne. L'obtention des fichiers XML de départ - c'est-à-dire le processus d'acquisition des caractères du dictionnaire papier et du balisage de formatage - peut être réalisé soit via encodage manuel, soit par scan puis OCR. La rétroconversion d'un article du FEW permet d'obtenir algorithmiquement un document XML au format FEW Semantic Markup Language (FSML).

2.2 Séquence de rétroconversion

Une trentaine d'algorithmes raffinent successivement le document XML représentant l'article à rétroconvertir. Le processus part d'un document XML au format FFML et ajoute des balises pour obtenir des documents XML au format FSML offrant une structuration de l'article de plus en plus complète. A chaque article est donc associé un ensemble de N (avec N ~ 35) documents XML: 1 document XML au format FFML, plusieurs dizaines de documents au format FSML (cf. figure 1).

Figure 1: Séquence de rétroconversion, du papier vers le balisage sémantique (FSML) en plusieurs étapes, en passant par un balisage de mise en forme (FFML)

2.3 Correction manuelle des articles partiellement rétroconvertis

Plusieurs sources d'erreurs peuvent entraver la séquence de rétroconversion. Tous les documents XML sont validés par rapport à des schémas XML (FFML ou FSML). Mais de nombreux problèmes subsistent: par exemple, une balise de formatage valide mais vide de sens, ou des erreurs dans le contenu textuel, tels que des artefacts introduits par le processus de numérisation. Les algorithmes de rétroconversion (identification et balisage sémantique) étant complexes, la séquence de rétroconversion échouera pour un certain nombre d'articles (nombre que l'on tente de minimiser par une analyse minutieuse de la structure profonde du FEW, cf. Büchi 1996, 117).

Par conséquent et essentiellement sur la base des indications de l'algorithme ayant échoué, des corrections manuelles devront être apportées à un certain nombre de documents XML: soit aux documents d'entrée, soit à des documents générés et partiellement rétroconvertis. Un exemple typique consiste à corriger un numéro d'appel de note de bas de page mal numérisé. Un autre exemple est la suppression de balises de formatage redondantes, ou encore le déplacement d'une balise sémantique mal positionnée par l'algorithme précédent, par exemple en raison de ponctuation erronée.

3 Plate-forme de rétroconversion distribuée

3.1 Modèle de rétroconversion distribuée

En considérant la complexité du problème, la masse de données à traiter et l'imperfection intrinsèque du processus de rétroconversion algorithmique, la rétroconversion du FEW nécessitera un certain temps et la contribution de nombreux experts. Inspirés par les projets Wikipedia (cf. Wikipedia), Gutenberg (cf. Projet Gutenberg) et Knol (cf. Knol), nous proposons un modèle de rétroconversion spatialement et temporellement distribuée, c'est-à-dire une édition coopérative et externalisée à large échelle des articles du FEW au fil des années.

(1) insertion d'un document FFML	si échec d'un algorithme:
ou	
(1) sélection d'un document FFML préalablement inséré	(3bis) téléchargement document FSML
(2) activation de la séquence de rétroconversion (documents FSML)	(4) correction manuelle (éditeur XML sur PC du contributeur)
(3) contrôle éditorial (documents FSML)	(5) insertion du document FSML modifié
	(6) retour en (2)

Figure 2: *Workflow* (organisation du travail) de 1 contributeur pour 1 article du FEW

Pour un article du FEW donné, l'organisation du travail des contributeurs du projet (experts du FEW et sous-traitants experts en acquisition de données papier) est résumée par la figure 2.

3.2 Design d'une plate-forme de rétroconversion distribuée

En theorie, il n'y a pas de différence entre la theorie et la pratique. En pratique, il y en a.
Yogi Berra

On peut considérer que les données rétroconverties et en cours de rétroconversion constituent un dépôt de documents (cf. figure 3). Au minimum, trois quarts de million de documents (~20000 articles x ~36 versions intermédiaires) entreront ou seront algorithmiquement générés dans le dépôt de documents, à quoi il faut ajouter les copies de documents qui en sortiront pour être lues, plus tous les documents réinsérés ou regénérés à la suite de corrections manuelles. Etant donné qu'il s'agit de données textuelles fortement redondantes et vu les capacités actuelles du matériel informatique, le stockage des données ne pose plus de problème significatif. Par contre, la gestion et la mise à disposition d'une telle quantité de données sont loin d'être triviales.

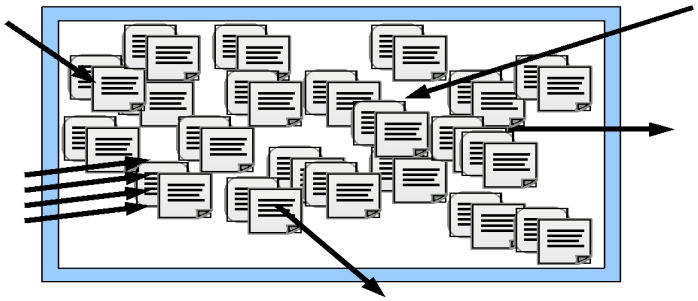

Figure 3: Dépôt de documents

Des solutions de gestion et de partage de données qu'on pourrait qualifier de classiques sont clairement non adaptées. L'échange direct de données via e-mail, clé USB, CD-ROM,... demanderait aux personnes gestionnaires des données une discipline probablement... inhumaine, ne garantirait pas la consistance des données (d'autant moins que le projet s'étendra vraisemblablement sur plusieurs années) et induirait des délais dans les flux de documents ne permettant pas une utilisation en temps réel. Le dépôt de documents doit donc être géré logiciellement. Une arborescence de dossiers partagés sur un serveur de fichiers n'offre pas de contrôle de versions de fichiers, offre un contrôle d'accès très basique et très difficilement accessible par plusieurs équipes. Un système de contrôle de version (Pilato 2004) lèverait toutes ces limites et permettrait théoriquement de rétablir la consistance du dépôt de documents après détection d'une erreur, mais est conçu pour un niveau d'utilisation plus proche du système que de l'utilisateur, notamment au niveau du contrôle d'accès.

Nous proposons donc une plate-forme de rétroconversion distribuée basée sur le web (Davidson 1999), c'est-à-dire accessible via un navigateur web (e.g. Chrome, Firefox, Internet Explorer, Opera, Safari) et ne nécessitant pas l'installation de logiciels par les utilisateurs (i.e. contributeurs). Cela permet une grande souplesse d'accès et d'utilisation.

Par exemple, une contributrice peut corriger un article à partir de son netbook via le navigateur Firefox et la connexion WiFi offerte par le café où elle tient une réunion informelle avec une collègue, et pourtant ses contributions sont immédiatement disponibles pour tous les autres contributeurs, sans qu'aucun d'entre eux n'ait à gérer la consistance et la synchronisation des données.

Concrètement, le dépôt de documents est géré par un serveur web applicatif s'appuyant, outre la séquence d'algorithmes de rétroconversion, sur un système de stockage de données (e.g. base de données relationnelle ou non) et un système de comptes utilisateurs. Nous proposons que les documents en cours de modification par un contributeur soient verrouillés. Tout ces éléments peuvent être mis en oeuvre de manière à garantir la consistance et synchronisation des documents XML (et également des algorithmes de rétroconversion en considérant qu'ils pourraient être mis à jour pour corriger des bugs). De plus, ces propriétés sont maintenues de manière totalement transparente pour les utilisateurs de la plate-forme.

Les opérations de base sur le dépôt de documents supportées par la plate-forme de rétroconversion distribuée sont les suivantes. La validation, l'insertion, la suppression et la génération de documents XML (i.e. activation de la séquence de rétroconversion) sont toutes réalisables en 1 ou 2 clics (incluant une confirmation). Le téléchargement et la visualisation de documents sont disponibles aux formats XML et HTML, respectivement. Une présentation relativement fidèle des articles est composée sur la base des documents XML et d'un fond d'écran beige. Un outil de mise en évidence des différences entre versions partiellement balisées d'un même article (i.e. de documents générés séquentiellement par la séquence de rétroconversion) est également disponible (cf. JMeld). Cela est très utile pour l'utilisateur qui veut vérifier les modifications apportées par les algorithmes de rétroconversion. Finalement, un moteur de recherche plein texte (donc non sémantique) et un index des articles sont déjà proposés dans le prototype logiciel implémenté. A terme, pourra venir s'ajouter un moteur de recherche sémantique.

3.3 Une architecture reposante

De manière à rendre reposants (i.e. faciliter) l'accès aux documents ainsi que l'activation des opérations qui peuvent leur être appliquées, nous proposons de construire la plate-forme de rétroconversion suivant le style architectural REST (REpresentational State Transfer) introduit par Roy Fielding dans sa thèse doctorale (Fielding 2000). L'architecture logicielle d'un système hypermédia distribué est considérée comme reposante (*RESTful*, cf. REST) si et seulement si toute requête d'accès ou modification de l'état courant du système est représentable par une requête HTTP et accessible par un hyperlien. Une conséquence immédiate est l'indépendance de chaque requête.

Tout document et toute opération sur n'importe quel document sont accessibles en un clic. Par exemple, la séquence de rétroconversion est appliquée au document identifié par fewa-id-1234 (préalablement inséré dans le dépôt de documents) en visitant l'URL[1]: http://www.few-prototype.info/workspace/retroconversion/fewa-id-1234/post

1 En supposant que la plate-forme de rétroconversion soit déployée sur le domaine fictif www.few-prototype.info

Les documents générés sont aussi immédiatement disponibles via leur propre URL.
Le recours à une plate-forme web apporte la transparence du maintien de la consistance et de la synchronisation du système vis-à-vis de ses utilisateurs (i.e. contributeurs). L'architecturer suivant le style REST permet une très grande souplesse à la fois d'accès et d'implémentation. Il n'est en effet pas nécessaire de spécifier quels composants logiciels seront sélectionnés pour traiter les requêtes, servir et stocker les données. Il suffit de décrire les interactions avec le système (i.e. l'ensemble des URL et leur sémantique) pour décrire le comportement de celui-ci. La table 1 ci-dessous liste de manière exemplative les interactions nécessaires pour supporter les opérations sur le dépôt de document décrites à la section 3.2.

Visualisation de document	http://www.few-prototype.info/article/Mabille http://www.few-prototype.info/article/Mabille/5
Insertion de doucment Insertion de document partiellement rétroconverti	http://www.few-prototype.info/workspace/article-set http://www.few-prototype.info/... ...workspace/retroconversion/Mabille/post
Validation de document	http://www.few-prototype.info/workspace/validation
Activation de la séquence de rétroconversion	http://www.few-prototype.info/workspace/... ...retroconversion/fewa-id-000000001234/post
Recherche de document	http://www.few-prototype.info/search?a=mabillarde
Index des documents	http://www.few-prototype.info/index

Table 1: Exemples d'URL reposantes de la plate-forme de rétroconversion distribuée

4 Perspectives

Plusieurs extensions peuvent certainement être apportées à la plate-forme web reposante que nous avons introduite dans cet article, notamment au niveau de l'organisation des contributeurs (cf. figure 2) et du modèle de verrouillage des articles (cf. section 3.2).

Ainsi, serait-il intéressant d'introduire un modèle de verrouillage plus fin, par exemple au niveau du paragraphe plutôt que de l'article? Ou bien de remplacer le modèle de verrouillage par un modèle de réconciliation de version, e.g. algorithme *3-way merge* (Lindholm 2003)? Serait-il en outre intéressant de supporter le maintien de plusieurs versions rétroconverties d'un même article, résultant de l'application de séquences de rétroconversion différentes et/ou de corrections manuelles différentes? Dans quelle mesure un outil de suivi de bugs, e.g. *bug tracker* (Gutwin 2004) peut-il faciliter les discussions entre contributeurs au sujet de documents à corriger? Enfin, quel est le niveau de détails souhaitable pour la *provenance* (Simmhan 2005) des documents?

Clairement, la plate-forme de rétroconversion que nous proposons est taillée sur mesure pour être déployée sur un *cloud* (Buyya 2008). En même temps, il sera utile d'introduire un modèle hybride d'utilisation connectée/non-connectée. Le support de la mise en cache locale de documents dans le navigateur web (cf. Gears, cf. HTML 5) permettra, dans une certaine mesure, l'utilisation du système en l'absence de connectivité Internet.

Remerciements

Nous tenons à remercier les participants de la journée TraSoGal pour les discussions qui ont permis d'enrichir notre réflexion; Dr. Esther Baiwir, France Gabriel et Marie Steffens pour leur évaluation d'une version de test de la plate-forme de rétroconversion; Kees Kuip (auteur du logiciel de diff' visuel JMeld); ainsi que Xavier Dalem pour ses suggestions très pertinentes à propos de l'interface web du système.

Références

Büchi, Éva, 1996. «Les Structures du /Französisches Etymologisches Wörterbuch/. Recherches métalexicographiques et métalexicologiques», Tübingen, Niemeyer.

Buyya, Rajkumar, Yeo, Chee Shin, Venugopal, Srikumar, 2008. «Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities», Keynote Paper, Proceedings of the IEEE International Conference on High Performance Computing and Communications, Dalian, China.

Davidson, James Duncan et Coward, Danny, 1999. «Java Servlet Specification», Sun Microsystems.

FEW = Wartburg (Walther von) et al., 1922–2002. Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes, 25 volumes, Bonn/Heidelberg/Leipzig-Berlin/Bâle, Klopp/Winter/Teubner/Zbinden.

Fielding, Roy, 2000. *Architectural Styles and the Design of Network-based Software Architectures*, Ph.D. dissertation, UC Irvine.

Google Gears, dernière consultation le 24 septembre 2009, http://gears.google.com/

Gutwin, Carl, Penner, Reagan et Schneider, Kevin, 2004. «Group awareness in distributed software development», Proceedings of the ACM Conference on Computer Supported Cooperative Work, Chicago, IL, USA .

HTML 5, dernière consultation le 24 septembre 2009, http://www.w3.org/html/wg/html5/

JMeld, dernière consultation le 24 septembre 2009, http://www.xs4all.nl/~keeskuip/jmeld/

Knol, dernière consultation le 24 septembre 2009, http://knol.google.com/k?hl=fr

Lindholm, Tancred, 2003. «XML three-way merge as a reconciliation engine for mobile data», Proceedings of the ACM International Workshop on Data Engineering for Wireless and Mobile Access, San Diego, CA, USA.

OED = Oxford English Dictionary, dernière consultation le 24 septembre 2009, http://www.oed.com/

Pilato, Michael, 2004. *Version Control With Subversion*, O'Reilly.

Projet Gutenberg, dernière consultation le 24 septembre 2009, http://www.gutenberg.org/

REST, article Wikipedia sur Representational State Transfer, dernière consultation le 24 septembre 2009, http://en.wikipedia.org/wiki/Representational_State_Transfer

Simmhan, Yogesh L., Plale, Beth et Gannon, Dennis, 2005. «A survey of data provenance in e-Science», ACM SIGMOD Record, Vol. 34, No. 3.

TLFi = CNRS/Université Nancy2/ATILF, 2004. *Trésor de la Langue Française informatisé* (cédérom), Paris, CNRS Editions (version web: http://atilf.atilf.fr/tlf.htm, dernière consultation le 24 septembre 2009).

Wikipedia, dernière consultation le 24 septembre 2009, http://www.wikipedia.org/