Xavier-Laurent SALVADOR. XML pour les linguistes. *L'Harmattan*. 2016. 189 pages. ISBN 978-2-34309-956-9.

Lu par **Lydia-Mai Ho-Dac**

Université de Toulouse – CLLE-ERSS

« XML pour les linguistes » *se donne pour objectif de présenter le langage XML et ses applications à des chercheurs en lettres et sciences humaines et sociales à même de produire des ressources langagières dans un format moins explicite et partagé que l'XML. L'objectif est clairement de rassurer un public parfois frileux devant la technologie informatique et de le convaincre que le langage XML est adapté aux recherches en linguistique. La présentation commence par les éléments historiques dont hérite le langage XML pour introduire de façon quelque peu littéraire le fonctionnement du langage XML. Viennent ensuite des parties plus didactiques qui expliquent la syntaxe XML, l'utilité et la création de DTD et schémas XSD, les normes XML (TEI, RDF, OWL) et les langages de requête et de transformation disponibles (Xpath, XSL, XQuery). La dernière partie présente trois projets menés par l'auteur mettant en jeu des ressources XML variées (corpus littéraires, dictionnaires, transcriptions) et l'outil Isilex développé par l'auteur.*

Cet ouvrage se donne pour objectif de présenter le langage XML et son potentiel à des étudiants et chercheurs en lettres et sciences humaines et sociales. Tout en définissant les règles de base du langage XML, l'auteur justifie, documente et illustre l'utilisation de cette norme pour des études en linguistique et plus largement des études manipulant du matériau langagier.

Les notions informatiques présentées sont régulièrement reliées à des notions historiques (le chevron, caractère utilisé par les copistes bibliques pour distinguer des éléments textuels différents, le codex comme éternelle base de représentation d'un texte, etc.) dans un souci de montrer que finalement rien n'est vraiment nouveau. Cette attention semble directement adressée à un public frileux à l'égard des technologies informatiques et que l'auteur souhaite convaincre de la simplicité et du bien-fondé du langage XML.

Les avantages du langage XML pour la recherche en lettres et sciences humaines et sociales sont documentés et illustrés à travers une variété d'exemples de ressources en XML (dictionnaires, lexiques, corpus de textes bruts et annotés, etc.) et de projets collaboratifs.

Organisation et contenu

L'ouvrage commence par une introduction à certaines notions fondamentales pour manipuler des formats numériques (le « plasma numérique » selon l'auteur) : octets, caractères et encodage, documents numérisés *vs* électroniques, formats de fichiers, etc.

Le chapitre 2 décrit le langage XML et un certain nombre de normes associées. Il commence par un aperçu des origines du langage XML en tant que convention typographique, puis de la syntaxe XML présentée comme permettant de décrire et déclarer des objets langagiers, à la manière des didascalies (« ceci est un paragraphe »). Après une définition des espaces de nommage, décrits comme des dialectes de XML, une large partie de chapitre 2 est dédiée à la nécessité de normaliser la description et l'encodage des éléments d'une ressource, ainsi qu'aux moyens utilisés pour assurer cette normalisation et, à terme, une pérennisation de la ressource. Sont parcourus les formalismes XML utilisés par les outils de traitement de texte (Open Office et MSWord), la gestion et la création de DTD et de schémas XML, la norme TEI pour l'encodage de ressources textuelles, le modèle RDF pour la structuration des données du Web sémantique, et le format OWL pour l'encodage de ressources terminologiques et les ontologies du Web.

Le troisième chapitre est consacré aux langages de requête et de transformation XPath, XSL et XQuery. La présentation est orientée de façon à montrer toute la valeur ajoutée d'une ressource structurée en langage XML. Le chapitre commence par les bases de la syntaxe d'une requête XPath et d'une transformation XSLT, avant de décrire le langage de requête et de transformation XQuery, langage qui sera utilisé dans les applications et projets présentés dans la suite de l'ouvrage. Le choix pour le XQuery est justifié par son potentiel (c'est un langage de programmation en tant que tel) et la possibilité de l'utiliser *via* le logiciel BaseX, une application multiplateforme *open source* développée à l'université de Konstanz. La fin de ce chapitre illustre le potentiel de la combinaison XML, Xpath et XQuery par des « exemples de manipulations de corpus en XQuery » avec le logiciel BaseX : exploration de corpus bruts ou annotés, annotation de corpus et mise en place d'une interface en ligne pour la consultation et l'interrogation d'une ressource XML. Le chapitre fini sur une liste éclectique d'outils permettant la création, la manipulation et/ou l'exploitation de ressources XML : Oxygen. ToolBox, TXM, Transcriber, PRAAT, ELAN, et Isilex, développé par l'auteur et utilisé dans les projets décrits dans le chapitre 4.

Le quatrième et dernier chapitre présente trois projets menés par X.-L. Salavador avec le logiciel Isilex. Ces projets illustrent des cas de (1) détection automatique de thématiques dans des textes littéraires ; (2) de visualisation d'annotations sous forme de graphes ; (3) de construction et gestion collaborative du dictionnaire Crealscience, dictionnaire de français scientifique médiéval ; (4) de transformation du format XML vers le format LaTeX ; (5) de transcription collaborative pour une version numérique et consultable de l'Exode de la *Bible historiale*.

Commentaire

« *XML pour les linguistes* » est conçu comme un livre qui se lit plus qu'un manuel qui s'utilise. L'auteur met l'accent sur le fait que le langage XML hérite de concepts issus de la tradition de l'édition et de l'écriture depuis ses origines, parfois au détriment d'indications simples pour la bonne prise en main du langage XML par un public néophyte.

Le caractère didactique est toutefois présent dans les trois premiers chapitres, notamment grâce à un lexique qui fournit des définitions complètes des termes utilisés dans l'ouvrage et également à un certain nombre d'encadrés offrant des résumés didactiques (principales règles de syntaxe XML, procédures pour construire un projet collaboratif de base de données XML, etc.) et des mini-tutoriels (gérer des fichiers XML en ligne de commande, etc.). On regrettera cependant le peu d'outils pour faciliter la navigation à l'intérieur du document : pas de renvoi depuis le lexique aux pages pertinentes, absence de table des encadrés, des figures et exemples utilisés, faible nombre de titres de section pour pointer sur des aspects spécifiques du langage XML.

Cet ouvrage s'adresse donc à des étudiants et chercheurs en lettres et sciences humaines et sociales qui ne cherchent pas un manuel, mais plutôt une sorte de cours à lire pour comprendre pourquoi et comment utiliser le langage XML. La première partie peut tout à fait être utilisée en support de cours par des néophytes. La partie consacrée à la manipulation de ressources XML par le langage XQuery et le logiciel BaseX est davantage construite comme un tutoriel adressé à un public plus averti ayant certaines compétences en programmation et gestion de systèmes informatiques, des « ingénieurs du texte » selon les propos de l'auteur. Le dernier chapitre sert, quant à lui, à présenter des projets menés par l'auteur, sans spécialement adopter une orientation didactique.