briefing paper

digital preservation europe

L'extraction automatisée des métadonnées sémantiques

Depuis mai 2005, the Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, a démarré le projet d'automatiser l'extraction des métadonnées sémantiques des objets numériques. Ce projet s'appuie sur une étude précédente sur l'automatisation ou la semi automatisation des procédures d'entrée et de préservation (par exemple [2]). La constitution des métadonnées nécessaires pour décrire le contenu, les informations bibliographiques, la provenance, les besoins techniques et administratifs d'un objet, est un élément crucial de gestion et de subsistance des dépôts numériques, des bibliothèques et des centres d'archives ([8], [9]). L'élaboration manuelle de telles métadonnées est un labeur intensif, et, à la vitesse exponentielle à laquelle les objets numériques sont produits, il ne sera bientôt plus possible de ne se reposer que sur des méthodes manuelles. L'objectif de ce projet est de procéder par étape pour appréhender jusqu'où la création de telles métadonnées peut être automatisée, avant que l'urgence n'apparaisse.

Le champ d'action

Les efforts initiaux sont concentrés sur l'extraction des métadonnées descriptives des objets en format PDF : en sélectionnant un format spécifique, on peut réduire l'ampleur du problème à une taille gérable. Plus spécifiquement, un outil qui peut appréhender PDF, un format largement adopté par les dépôts numériques, les bibliothèques et les centres d'archives ainsi que le secteur commercial et les particuliers, est supposé être d'un usage immédiat pour un large spectre de communautés.

Les efforts initiaux sont concentrés sur les documents texte : les méthodes de processus en langage naturel (NLP) se sont révélées efficaces en ce qui concerne la recherche, l'extraction et la classification des documents et de leurs termes. Ceci pose NLP et les autres techniques automatisées d'apprentissage pour les documents texte en candidat évident pour remplir les premières étapes d'extraction des métadonnées. Le développement d'outils d'extraction pour le texte est aussi supposé avoir des conséquences sur les autres objets, car beaucoup de procédures d'extraction pour d'autres médias (image ou matériaux audiovisuels) dépendent de l'exploration des textes associés.

Le déroulement

A HATII, la classification automatique par genre a été retenue comme point fondamental pour réaliser l'extraction automatique des métadonnées. Le genre est une classification structurelle et fonctionnelle d'un document qui est le reflet de l'un des points suivants :

- L'intention du créateur (informer, argumenter, instruire),
- L'interprétation des utilisateurs (un ensemble de faits, une expression d'opinion, un article de recherche),
- La description d'un processus (article pour publication dans une revue, curriculum vitae, compte rendu de réunion)
- Le type de structure des données (table, graphique, carte, liste).

Pourquoi la classification par genre?

1. Identifier le genre limitera les possibilités structurelles des types de documents à partir desquels extraire les autres métadonnées :

- L'espace de recherches pour les autres métadonnées sera réduit ; pour un même genre, les métadonnées telles que l'auteur, les mots clés, les numéros d'identification ou les références devraient apparaître dans des styles et des localisations similaires.

2. Identifier le genre créera un outil dédié qui homogénéisera le travail spécifique au genre :

- Il existe des travaux indépendants ([1], [3], [4], [10], [11]) pour l'extraction de métadonnées d'un genre spécifique, qui peuvent être incorporés à un classificateur général de genres pour l'extraction des métadonnées dans beaucoup de domaines.
- Les ressources disponibles pour l'extraction d'autres métadonnées sont différentes d'un genre à l'autre ; par exemples, les articles de chercheurs, à l'inverse des articles de journaux, comportent une liste de références d'articles étroitement liés au document d'origine et conduisent ainsi à une meilleure classification.

3. Expérimenter des genres nouveaux qui n'apparaissent pas dans le contexte des bibliothèques conventionnelles est une nécessité afin de mettre à niveau les procédures de gestion des matériaux numériques.
4. Différentes politiques institutionnelles de collecte pourraient être concentrées sur divers genres de matériaux numériques. La classification par genre aidera l'automatisation de l'identification, de la sélection et de l'acquisition de ces matériaux tout en respectant les lignes directrices des collectes locales.

L'outil de classification

L'outil de classification des genres est destiné à évaluer des documents statistiquement basés sur leur aspect visuel (par exemple la quantité d'espace inoccupée dans un document), leurs éléments de style (la fréquence et l'utilisation de certains articles), les modèles de langue (les termes significatifs qui caractérisent le document), les formes sémantiques (l'utilisation de phrases avec des noms subjectifs) et les ressources extérieures (les sources URL), afin de déterminer leur classe de genre. Les résultats de ces travaux de recherches ont été publiés dans plusieurs articles ([5], [6], [7]).

La procédure d'extraction des métadonnées

L'étude dans ce projet sera dévolue à l'architecture générale d'une procédure d'extraction et d'ingestion automatisées des métadonnées selon le processus suivant :

1. Recevoir l'objet numérique,
2. Déterminer son genre ou sa classe structurelle,
3. Evaluer le meilleur outil d'extraction des métadonnées pour le genre identifié ou la meilleur option basée sur la structure,
4. Si un tel outil n'existe pas, en demander la création,
5. Une fois l'outil choisi ou créé, extraire les métadonnées requises,
6. Enregistrer l'objet et ses métadonnées dans le dépôt ou le centre d'archives.

Conclusions

Les processus automatisés d'enregistrement, de préservation et de sélection dans un dépôt numérique ne sont plus une facilité mais une nécessité. Le stockage, le partage et l'utilisation inter institutions de l'information sont maintenant devenus une réalité active ; le contrôle manuel de tels processus est lent, inefficace et insuffisant. Les défis d'adaptation aux autoroutes de l'information doivent être relevés par des processus automatisés et innovants d'extraction, d'authentification et d'évaluation. L'extraction automatisée des métadonnées sémantiques n'en est encore qu'à ses balbutiements. Il est essentiel que les efforts soient poursuivis afin de mettre en avant et de peaufiner des outils d'extraction et de les intégrer aux autres procédures dans les bibliothèques, les centres d'archives et autres communautés concernées.

Références

[1] Bekkerman, R., McCallum, A., Huang, G. (2004) Automatic Categorization of Email into Folders. Benchmark Experiments on Enron and SRI Corpora', CIIR Technical Report, IR-418.

[2] ERPANET: Packaged Object Ingest Project. http://www.erpanet.org/events/2003/rome/presentations/ross_rusbridge_pres.pdf

[3] Han, H., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E. A. (2000) Automatic Document Metadata Extraction using Support Vector Machines. Proceedings 3rd ACM/IEEECS Conference on Digital libraries, 37-48.

[4] Giurida, G., Shek, E. Yang, J. (2000) Knowledge-based Metadata Extraction from PostScript File. Proceedings 5th ACM International Conference on Digital Libraries, 77-84.

[5] Kim, Y. and Ross, S. (2007) Detecting family resemblance: Automated genre classification. CODATA Data Science Journal, Volume 6, S172-S183, ISSN:1683-1470.

[6] Kim, Y. and Ross, S. (2006) Genre classification in automated ingest and appraisal metadata. In J. Gonzalo, editor, Proceedings European Conference on advanced technology and research in Digital Libraries (ECDL), volume 4172 of Lecture Notes in Computer Science, pages 63–74. Springer.

[7] Kim Y. and Ross, S. (2006) "The Naming of Cats": Automated genre classification. To appear International Journal of Digital Curation, preprint available at http://eprints.erpanet.org/123

[8] PREMIS (PREservation Metadata: ImplementationStrategy) Working Group: http://www.oclc.org/research/projects/pmwg/

[9] Ross S and Hedstrom M. (2005) Preservation Research and Sustainable Digital Libraries. International Journal of Digital Libraries (Springer) DOI: 10.1007/s00799-004-0099-3.

Formatted: Bullets and Numbering

Auteur:Yunhyong Kim, *DCC-EPSRC Research Fellow, HATII University of Glasgow*