PRÉSERVER LA DIVERSITÉ DES CONNAISSANCES DANS UN CONTEXTE MULTILINGUE ET MULTICULTUREL

ENJEUX ET DÉFIS POUR LES SCIENCES DE L'INFORMATION

PRESERVING KNOWLEDGE DIVERSITY IN MULTILINGUAL AND MULTICULTURAL CONTEXT

ISSUES AND CHALLENGES FOR INFORMATION SCIENCE

JOURNÉE D'ÉTUDE ORGANISÉE PAR / ONE-DAY CONFERENCE ORGANIZED BY

AMEL FRAISSE

LABORATOIRE GERIICO, AXE 4, UNIVERSITÉ DE LILLE

DATE: JEUDI 17 JUIN 2021 / THURSDAY JUNE 17TH 2021

10 AM – 4 PM (LOCAL TIME : UTC+2H)

SITE WEB / WEBSITE:

HTTPS://KNOWLEDGE-DIVERSITY.UNIV-LILLE.FR

EN LIGNE VIA ZOOM / ONLINE VIA ZOOM

LIEN POUR PARTICIPER / LINK TO PARTICIPATE :

HTTPS://UNIV-LILLE-FR.ZOOM.US/J/97835811279?PWD=DUQVQJZESUZOY2PVMGXANEPTSXH4DZ09

ID DE REUNION / ID ZOOM : 978 3581 1279

Géricico

Université de Lille

INTERVENANTS / SPEAKERS

10:30 AM	DAMIEN NOUVEL
11:15 AM	FANNY MION-MOUTON
2:30 PM	VICTORIA VAN HYNING
3:00 PM	BEN W. BRUMFIELD & SARA BRUMFIELD

PROGRAMME / PROGRAM

10:00 – 10:30 AM SESSION D'OUVERTURE / OPENING SESSION

10:00 AM Ouverture de la journée

WIDAD MUSTAFA EL HADI, GERIICO, UNIVERSITE DE LILLE

10:15 AM Introduction

AMEL FRAISSE, GERIICO, UNIVERSITÉ DE LILLE

10:30 – 12:00 AM PANEL 1

MODÉRÉ PAR / MODERATED BY JOANA CASENAVE

10:30 AM Doter les langues en ressources numériques (données et outils) : Un retour d'expérience de projets à l'INALCO

DAMIEN NOUVEL, INSTITUT NATIONAL DES LANGUES ET CIVILISATIONS ORIENTALES, PARIS

11 :15 AM 350 langues, 80 alphabets : Multilinguisme et multiculturalisme en bibliothèque, à travers l'exemple de la BULAC

FANNY MION-MOUTON, BIBLIOTHEQUES UNIVERSITAIRE DES LANGUES ET CIVILISATIONS, PARIS

2 :30 – 3 :30 PM PANEL 2

MODÉRÉ PAR / MODERATED BY RONALD JENN

2 :30 PM Multilingual collections, crowdsourcing, and staff expertise: unexpected upsides of the global pandemic

VICTORIA VAN HYNING, UNIVERSITY OF MARYLAND, COLLEGE OF INFORMATION STUDIES (ISCHOOL), USA

3:00 PM Lessons from 5 years of indigenous language transcription projects

BEN W. BRUMFIELD & SARA BRUMFIELD, BRUMFIELD LABS, USA

3:30 – 4:00 PM DISCUSSION GENERALE / GENERAL DISCUSSION

MODEREE PAR / MODERATED BY AMEL FRAISSE

Gériico

Université de Lille

SHORT BIOS

DAMIEN NOUVEL

(FR) Damien NOUVEL est maître de conférences en informatique à l'Institut National des Langues et Civilisations Orientales (INALCO) au sein du laboratoire ERTIM, et directeur de cette équipe depuis 2020. Il travaille dans le domaine du traitement automatique des langues (TAL) avec un intérêt plus particulier pour les modèles mathématiques et le multilinguisme, dont quelques langues enseignées à l'Inalco (quechua, arabe, bambara, chinois, etc.) pour des objectifs variés (translittération, désambiguisation lexicale, opinion, analyse textométrique, etc.).

(EN) Damien Nouvel is an associate professor at INALCO within ERTIM research team, and head of this team since 2020. His research focus on Natural Language Processing (NLP) and mixes mathematics (statistics, machine learning), computer sciences (resources and tools) and linguistics (with a special interest in low-resourced languages, for instance lately quechua or bambara).

FANNY MION-MOUTON

(FR) Fanny Mion-Mouton est responsable adjointe du pôle flux et données et responsable de l'équipe signalement et exposition des données à la BULAC. Archiviste paléographe de formation, elle a suivi la formation des conservateurs de bibliothèques à l'ENSSIB avant de rejoindre la BULAC en juillet 2013. Dans le cadre de ses fonctions, elle participe à différents projets liés à l'informatique documentaire (SIGB Koha) ou la gestion de la numérisation.

(EN) Fanny Mion-Mouton is deputy head of the flow and data division and head of the reporting and data exposure team at BULAC. A paleographer archivist by training, she followed the training of library curators at ENSSIB before joining BULAC in July 2013. As part of her duties, she participates in various projects related to documentary computer science (SIGB Koha) as well as digitization management.

VICOTRIA VAN HYNING

Victoria joined the iSchool in 2020 and is an affiliate of the English Department. From 2018-2020 she served as a Senior Innovation Specialist for the Library of Congress' crowdsourcing project By the People. She held a British Academy Postdoctoral Fellowship in English literature at Oxford University, where she also served as the Humanities PI of the crowdsourcing program Zooniverse.org (2015-2018). Her teaching and research interests focus on giving more oxygen to marginalized voices and people, whether in the historical record—such as religious minorities, women, and Black artists—or people alive today. She leads the David C. Driskell Papers Project with Driskell Center colleagues, and is a founding member of the Center for Archival Futures (CAFe), and Data Rescue and Reuse (RRAD) Lab where she leads investigations about the long-term preservation, use and reuse of crowdsourced data. She has emerging interests in prison librarianship and education, and supporting returning citizens.

BEN W. BRUMFIELD

Ben Brumfield is a partner at FromThePage, a collaborative platform for transcribing, translating, and indexing manuscripts. After early experiences editing Wikipedia and Pepys Diary Online, he was inspired to build crowdsourcing software to transcribe a set of family diaries in 2005 and released it as an open-source tool in 2009. FromThePage has since been adopted by libraries, archives, and researchers for material ranging from financial records to Aztec codices. He writes about crowdsourcing and textual encoding on the FromThePage project blog.

SARA BRUMFIELD

Sara Brumfield is a partner at FromThePage, where she builds software and helps state and national archives, research groups, public libraries, and universities run crowdsourcing projects. Prior to FromThePage, Sara spent 17 years as a software engineer with IBM. She led development and support teams focused on system and network management products, serving as a focal point for large enterprise customers. She holds eight technical patents. She has a BA in Computer Science and the Study of Women and Gender from Rice University.

ABSRTACTS

DOTER LES LANGUES EN RESSOURCES NUMERIQUES (DONNEES ET OUTILS) : UN RETOUR D'EXPERIENCE DE PROJETS A L'INALCO

DAMIEN NOUVEL

Multilingual considerations are much related to languages vitality, which raises several challenges, including technological issues that are undeniably a major concern, related to both Computer Sciences, Linguistics and Natural Language Processing. Challenges in this regard have evolved over the last decades. Raw data needs have continuously increased, both from written (texts) and oral (audio) sources. Robust automatic processings (OCR/ASR) now provide useful tools that can today be used and improved (learned) by non-NLP users, corresponding transcription and annotation efforts are thus reduced. Low level linguistic annotation (i.e. segmentation, POS) needs are still present, but are easier to implement using supervised (embeddings) or unsupervised methods and/or transfer learning. Syntax, semantic and understanding tasks are still major challenges. In this talk, I will present past and ongoing projects raising those questions and a more general overview of current trends, both in academics or contributive communities.

350 LANGUES, 80 ALPHABETS : MULTILINGUISME ET MULTICULTURALISME EN BIBLIOTHEQUE, A TRAVERS L'EXEMPLE DE LA BULAC

FANNY MION MOUTON

La Bibliothèque universitaire des langues et civilisations, ouverte en décembre 2011, concentre dans ses murs des collections qui concernent l'ensemble des civilisations et langues du monde non occidental : 1,5 million de documents, 350 langues et 80 écritures. La particularité de ces fonds, d'une extrême variété, influe sur l'organisation du travail et induit des problématiques spécifiques, tant du point de vue de l'acquisition de la documentation, que de son traitement. Le développement et la communication des collections sur les domaines couverts par la BULAC supposent une adaptation des cadres habituellement utilisés en bibliothèque. L'acquisition de la documentation multilingue, qu'elle soit papier ou électronique, implique la mise en place de circuits adaptés à chaque fonds, l'insertion de l'établissement dans des réseaux et consortiums européens et internationaux aréaux, ou encore l'utilisation d'une classification spécifique. Le traitement catalographique de ces collections soulève également des enjeux particuliers, pour permettre la cohabitation de données multilingues et multi-écritures au sein d'un même catalogue, tout en s'insérant dans le réseau national de catalogage de l'Agence bibliographique de l'Enseignement supérieur (Sudoc). Dans le cadre de la mise en place de la Transition bibliographique, la qualité des données et leur interopérabilité sont autant d'enjeux majeurs, sur lesquels la BULAC s'efforce d'avancer, en tenant compte de ses spécificités. Ces dernières années, des projets importants ont ainsi été menés sur l'amélioration de l'indexation du catalogue, l'enrichissement et l'alignement des données, ou encore la réflexion sur les normes de translittération, afin de valoriser et d'exposer autant que possible une documentation riche et diversifiée.

MULTILINGUAL COLLECTIONS, CROWDSOURCING, AND STAFF EXPERTISE: UNEXPECTED UPSIDES OF THE GLOBAL PANDEMIC

VICTORIA VAN HYNING

When large collections of materials such as presidential papers, the papers of society elites, merchants or activists are curated at large institutions such as the Library of Congress, the British Library or Bibliotèque Nationale, the scale of the collections makes it nearly impossible to capture item-level linguistic information or other details such as named entities, the presence of images, material type or indeed the text of the documents themselves. These challenges are present at most institutions of all sizes. Crowdsourcing is an invaluable tool for gathering many of these types of information not traditionally identified by staff, but not all "crowds" will include people with relevant specialist language knowledge to accurately transcribe or translate the languages in a collection. In this talk I will give an overview of a large crowdsourcing transcription project By the People (crowd.loc.gov) at the Library of Congress (LOC), in which members of the public are invited to

transcribe and edit one another's transcriptions of LOC materials. These transcriptions are published on the Library's website (loc.gov) and make the collections both more discoverable to researchers and accessible for people who use screen readers. The project launched in October 2018 and by February 2020 volunteers had completed transcribed and reviewed 50,000 pages. By May the completed page count was over 100,000, and 200,000 by late August. These exponential increases were partly driven by a spike in media coverage of crowdsourcing projects all over the world, which encouraged people to partake in crowdsourcing while locked down to keep loneliness at bay, enable connection with other virtual volunteers as well as culture, keep their minds active or distracted from the news, learn a new skill, and more. It was also driven by the participation of LOC staff, who brought their specialist language and other skills to bear on the collections. This paper will discuss these unexpected upsides of the pandemic, and offer suggestions for longer-term uses of crowdsourcing to surface specialist materials from heterogeneous collections.

LESSONS FROM 5 YEARS OF INDIGENOUS LANGUAGE TRANSCRIPTION PROJECTS

BEN W. BRUMFIELD & SARA BRUMFIELD

Over the last five years we've hosted transcription projects in non-dominant or indigenous languages including Nahuatl, Mixtec, Dakȟóta/Lakȟóta, Diyari, Jawi, Old French, Old English, Latin, Dutch, and Arabic. We've learned that multilingual transcription isn't simple and indigenous communities have specialized needs. We bring tool-maker's perspective to the technical and collaborative challenges these projects present.