

POWER AND ETHICS OF ARTIFICIAL INTELLIGENCE FOR MINING AND EXPLORING HERITAGE AND BIBLIOGRAPHICAL DATA

POUVOIR ET ÉTHIQUE DE L'INTELLIGENCE ARTIFICIELLE POUR LA FOUILLE ET L'EXPLORATION DES DONNÉES PATRIMONIALES ET BIBLIOGRAPHIQUES

Mercredi 21 juin : 14h-17h30
(en mode hybride)

Salle Gérard Losfeld (B2.454)
Campus Pont de Bois

Lien zoom :

<https://univ-lille-fr.zoom.us/j/96757776980?pwd=b3lqWHBLQzBreGI2eDRmVXh5MjBsUT09>
ID de réunion: 967 5777 6980

Séminaire de Recherche organisé par

Amel Fraisse

iSchool, Université de Lille, Gériico
amel.fraisse@univ-lille.fr

Axe 4 : Organisation des Connaissances et Circulation de l'Information

PROGRAMME

14h-14h30

Introduction

**Amel FRAISSE, Maitresse de conférence, Université de Lille, Laboratoire
GERiCO**

14h30-15h30

Holistic Semantic Representation and Handling of Library Information

Stefano Ferili, Associate Professor at University of Bari, Italie

15h30-16h30

Les collections numériques à la BnF : services et perspectives de recherche

Marie Carlin, coordinatrice du BnF DataLab, Paris

16h30-17h30

Les enjeux de l'IA vu par des infrastructures de recherche / Artificial Intelligence as seen by
European and National Research Infrastructures

**Edward Gray : Chargé de recherche, TGIR HUMA-NUM, CNRS, Paris
& Camille Ferrari : Bibliothèque apostolique vaticane**

ABSTRACTS/ RÉSUMÉS

ABSTRACT: Holistic Semantic Representation and Handling of Library Information

The traditional record-based approach used in library practice has gone as far as it could. With the advent of Digital Libraries, new models oriented toward relational representations have been proposed. To support the needs and activities of different kinds of users, we push forward this perspective and propose a graph-based 'holistic' representation of DL knowledge, describing the documents' metadata, physical aspects, content, context and lifecycle/usage, beyond the traditional. We also propose a framework based on Formal Ontologies and LPG-based DBs that may overcome some limitations of the traditional Semantic Web approach based on the RDF model, and enables a wider range of automated reasoning approaches. Connected issues are Document Image Analysis and Understanding and Knowledge Extraction (using NLP and image processing techniques). We developed prototypes of several systems covering the various aspects of this vision.

RESUME : Représentation sémantique holistique et traitement de l'information des bibliothèques

L'approche traditionnelle basée sur les enregistrements utilisée dans la pratique bibliothéconomique est allée aussi loin qu'elle le pouvait. Avec l'avènement des bibliothèques numériques, de nouveaux modèles orientés vers les représentations relationnelles ont été proposés. Pour répondre aux besoins et aux activités de différents types d'utilisateurs, nous poussons plus loin cette perspective et proposons une représentation "holistique" de la connaissance des bibliothèques numériques basée sur les graphes, décrivant les métadonnées, les aspects physiques, le contenu, le contexte et le cycle de vie/l'utilisation des documents, au-delà de l'approche traditionnelle. Nous proposons également un cadre basé sur des ontologies formelles et des bases de données basées sur les GPL qui peuvent surmonter certaines limitations de l'approche traditionnelle du web sémantique basée sur le modèle RDF, et permettre un plus large éventail d'approches de raisonnement automatisé.

Les questions connexes sont l'analyse et la compréhension d'images de documents et l'extraction de connaissances (à l'aide de techniques de NLP et de traitement d'images). Nous avons développé des prototypes de plusieurs systèmes couvrant les différents aspects de cette vision.

ABSTRACTS/ RÉSUMÉS

ABSTRACT : Digital collections at BnF: services and research perspectives

BnF's digital collections are extremely diverse, both in terms of form and content: catalog metadata, documents digitized on Gallica and GallicaIntramuros, databases, video games, Internet archives, connection logs, etc. Their recognition as a source and field for research raises issues of access and use, from a methodological, legal, and technical perspective. The requests for massive data extraction that have been multiplying over the past ten years have guided BnF's policy in terms of research support. What is the singularity of BnF data? What tools are used to exploit them? What are the terms of access and reuse? What are the limits to the opening of BnF data?

Through examples of research projects, this presentation will give an overview of the resources available and the search and retrieval tools offered by BnF, and will focus on the new research assistance service to be opened in October 2021, the BnF DataLab.

RESUME : Les collections numériques à la BnF : services et perspectives de recherche

Les collections numériques de la BnF sont d'une très grande diversité, tant par leur forme que leur contenu : métadonnées du catalogue, documents numérisés sur Gallica et GallicaIntramuros, bases de données, jeux vidéos, archives de l'Internet, logs de connexions... Leur reconnaissance comme source et terrain pour la recherche pose des problématiques d'accès et d'exploitation, sous un angle à la fois méthodologique, juridique et technique. Les demandes d'extraction massive de données qui se multiplient depuis une dizaine d'années ont orienté la politique de la BnF en matière d'accompagnement de la recherche. Quelle est la singularité des données de la BnF ? Quels outils permettent leur exploitation ? Quelles en sont les modalités d'accès et de réutilisation ? Quelles limites à l'ouverture des données de la BnF ?

A travers des exemples de projets de recherche, cette intervention propose de présenter un panorama des ressources disponibles et les outils de recherche et d'extraction proposés par la BnF, et de faire un focus sur le nouveau service d'aide à la recherche ouvert en octobre 2021, le BnF DataLab.

ABSTRACTS/ RÉSUMÉS

ABSTRACT : Artificial Intelligence as seen by European and National Research Infrastructures

Artificial Intelligence has profoundly changed the way research is conducted, and research.

Whether they be at the local, regional, national, or even international level, research infrastructures cannot ignore its impact on research practices and methodologies. The primary role of research infrastructures is to support and enable researchers, which usually regulates us to a second plan. However, it is necessary for research infrastructures to be leaders, especially on important fundamental issues that affect the field as a whole, such as artificial intelligence. Indeed, the ethical challenges posed by SSH research, and artificial intelligence in particular, should help guide the various actions of research infrastructure in our support to researchers.

This talk will show how the IR* HumaNum, the French national infrastructure for digital humanities, and DARIAH-EU, a European Research Infrastructure Consortium (ERIC) approach the question of artificial intelligence and how one must engage with this new field - as any research method of that matter - with ethics in mind.

RESUME : Les enjeux de l'IA vu par des infrastructures de recherche

L'intelligence artificielle a profondément modifié la façon dont la recherche est menée. Que ce soit à l'échelle locale, régionale, nationale ou même internationale, les infrastructures de recherche ne peuvent ignorer son impact sur les pratiques et les méthodologies de recherche.

Le rôle premier des infrastructures de recherche est de soutenir les chercheurs, ce qui nous ramène généralement à un second plan. Cependant, il est nécessaire que les infrastructures de recherche soient des leaders (force de proposition ? impulsent le mouvement, innovatrices...), en particulier sur des questions fondamentalement importantes qui affectent le domaine dans son ensemble, telles que l'intelligence artificielle. En effet, les défis éthiques posés par la recherche en sciences humaines, et l'intelligence artificielle en particulier, devraient guider les diverses actions des infrastructures de recherche dans leur soutien aux chercheurs. Cet exposé montrera comment l'IR* HumaNum, l'infrastructure nationale française pour les humanités numériques, et DARIAH-EU, un consortium européen d'infrastructures de recherche (ERIC) abordent la question de l'intelligence artificielle et comment il faut engager ce nouveau domaine, comme toute méthode de recherche d'ailleurs - en gardant l'éthique au premier plan.

ABSTRACTS/ RÉSUMÉS

ABSTRACT :

In addition to Edward Gray's presentation, Camille Ferrari, a student he supervises in the "Digital Technologies Applied to History" master's program at the École Nationale des Chartes, will speak about her internship, which uses artificial intelligence to help identify manuscript miniatures. She will present the internship she is currently doing at the Vatican Apostolic Library (BAV) in the coordination of IT services, under the supervision of Dr. Paola Manoni, Head of Coordination of IT Services and Dr. Giuliano Giuffrida. This presentation will present a concrete case of application of artificial intelligence on heritage collections. Indeed, she is working on the deployment of an artificial neural network aimed at identifying the elements of the miniatures of the manuscripts and the automatic creation of IIF annotations for each element, object, person or animal identified by the network. This project is made possible thanks to the use of an algorithm, named "You only look once" (YOLO), coded by BAV experts, and a data set designed for the segmentation and detection of objects on images called "Common Objects in Context" (COCO). Through a specific theme, animals, and with the scale of a sample of data, she wants to evaluate the relevance of the use of this neural network on humanist collections. Her intervention will focus on her first observations.

RESUME :

En complément de la présentation d'Edward Gray, interviendra Camille Ferrari, étudiante qu'il encadre au sein du master « Technologies numériques appliquées à l'histoire », à l'École nationale des chartes. Elle présentera le stage qu'elle effectue actuellement à la Bibliothèque apostolique vaticane (BAV), sous la direction du D r Paola Manoni, Responsable de la coordination des services informatiques et Giuliano Giuffrida, membre du même service et Docteur en astronomie. Cette présentation permettra d'exposer un cas concret d'application d'intelligence artificielle sur des collections patrimoniales. En effet, elle travaille sur le déploiement d'un réseau neuronal artificiel ayant pour but l'identification des éléments qui composent les miniatures des manuscrits et la création automatique d'annotations IIF pour chaque élément, objet, personne ou animal identifié par le réseau. Ce projet est rendu possible grâce à l'utilisation d'un algorithme, nommé « You only look once » (YOLO), experts de la BAV, et d'un set de données conçu pour la segmentation et la détection d'objets sur des images appelé « Common Objects in Context » (COCO). A travers une thématique spécifique, les animaux, et à l'échelle d'un échantillon de données, elle cherche à évaluer la pertinence de l'utilisation de ce réseau neuronal sur des collections humanistes. Son intervention se concentrera sur ses premières observations.

SHORT BIOS

Stefano Ferilli is an Associate Professor at the Department of Computer Science of the University of Bari, Italy, where from 2006 to 2018 he was the Director of the Inter-Departmental Center for Logic and Applications and since 2019 he leads the Automated Reasoning and Machine Learning laboratory. Since 2011 member of the Steering Committee of the Italian Association for Artificial Intelligence. Since 2019 member of the Steering Committee of the Italian Research Conference on Digital Libraries. His research interests are centered on Knowledge Representation and Multistrategy Reasoning. Applications include Digital Document Processing and Digital Libraries. (Co-)author of more than 340 National and International publications. Associate Editor of 2 international journals.

Stefano Ferilli est professeur associé au département d'informatique de l'université de Bari, en Italie, où, de 2006 à 2018, il a été directeur du centre interdépartemental pour la logique et les applications et, depuis 2019, il dirige le laboratoire de raisonnement automatisé et d'apprentissage automatique. Depuis 2011, il est membre du comité directeur de l'Association italienne pour l'intelligence artificielle. Depuis 2019, il est membre du comité directeur de la Conférence italienne de recherche sur les bibliothèques numériques. Ses recherches portent sur la représentation des connaissances et le raisonnement multistratégique. Les applications comprennent le traitement des documents numériques et les bibliothèques numériques. (Co-)auteur de plus de 340 publications nationales et internationales. Rédacteur en chef adjoint de deux revues internationales.

Marie Carlin is a library curator and coordinator of the BnF DataLab in the Department of Collection Discovery and Research Support.

A historian by training, she specializes in digital humanities and holds a master's degree in digital technologies applied to history from the École nationale des chartes.

For the past four years, she has been the coordinator of the BnF DataLab, a service that supports researchers who wish to conduct data mining on the BnF's collections.

Marie Carlin est conservatrice des bibliothèques et coordinatrice du BnF DataLab au sein du département de la Découverte des collections et de l'Accompagnement à la recherche

Historienne de formation, elle s'est spécialisée dans les Humanités numériques et est titulaire du master Technologies numériques appliquées à l'histoire de l'École nationale des chartes.

Depuis quatre ans, elle occupe le poste de coordinatrice du BnF DataLab, un service d'accompagnement des chercheurs qui souhaitent faire de la fouille de données sur les collections de la BnF

SHORT BIOS

Edward J. Gray is the Research Infrastructure Coordinator at the IR* Huma-Num (CNRS) and the Officer for National Coordination at DARIAH ERIC, the European Research Infrastructure for the Digital Arts and Humanities. He is currently on the Editorial Board for the SSH Open Marketplace, a discovery platform for digital humanities tools and services that was born from the SSHOC Project. He earned his doctorate in history from Purdue University, where his dissertation, "The Marillac: Family Strategy, Religion, and Diplomacy in the Making of the French State during the Sixteenth and Seventeenth Centuries," examined the ways in which familial politics impacted the formation of the early modern French state. While a doctorant invité at the École nationale des chartes in Paris, he earned a master's degree in Technologies numériques appliquées à l'histoire (TNAH), where he is also chargé de cours. He is also president of the ADEMEC, the Association des diplômé.es et étudiant.es et des Masters de l'École des chartes. Before working at Huma-Num, Edward was the Digital Humanities Coordinator at the Maison Européenne des Sciences de l'Homme et de la Société in Lille.

Edward J. Gray est chargé de recherche Infrastructure à l'IR* Huma-Num (CNRS) et Officer for National Coordination à DARIAH ERIC, l'infrastructure de recherche européenne pour les arts et les humanités. Il fait actuellement partie du comité éditorial du SSH Open Marketplace, une plateforme de découverte d'outils et de services pour les humanités numériques née du projet SSHOC. Il a obtenu son doctorat en histoire à Purdue University, où sa thèse, "The Marillac : Family Strategy, Religion, and Diplomacy in the Making of the French State during the Sixteenth and Seventeenth Centuries" (La stratégie familiale, la religion et la diplomatie dans la création de l'État français aux XVIe et XVIIe siècles), examine la manière dont la politique familiale a influencé la formation de l'État français au début de l'époque moderne. Doctorant invité à l'École nationale des chartes à Paris, il a obtenu un master en Technologies numériques appliquées à l'histoire (TNAH), où il est aujourd'hui chargé de cours. Il est par ailleurs président de l'ADEMEC, l'Association des diplômé.es et étudiant.es et des Masters de l'École des chartes. Avant de travailler à Huma-Num, Edward était coordinateur des humanités numériques à la Maison européenne des sciences de l'homme et de la société (MESHS) à Lille.

SHORT BIOS

Camille Ferrari is a master's student at the École nationale des chartes, in her second year of the "Digital technologies applied to history" degree. Holding two bachelor's degrees, in history and art history, and a master's degree in heritage and museums at the University of Paris 1 Panthéon-Sorbonne, she worked for the public service of the National Institute of History of Art, before joining the multimedia publishing department of the Bibliothèque nationale de France. She is currently completing her studies with an internship at the Vatican Apostolic Library, in the IT Services Coordination department, under the supervision of Dr. Paola Manoni, Head of Coordination of IT Services and Dr. Giuliano Giuffrida. Her work focuses on the deployment of an artificial intelligence on the manuscript collection of the Holy See.

Camille Ferrari est étudiante à l'École nationale des chartes au sein du master « Technologies numériques appliquées à l'histoire ». Diplômée de deux licences, en histoire et histoire de l'art, et d'une maîtrise en patrimoine et musées à l'université Paris 1 Panthéon-Sorbonne, elle a travaillé pour le service aux publics de l'Institut national d'histoire de l'art, avant de rejoindre le département des éditions multimédias de la Bibliothèque nationale de France. Elle achève actuellement son cursus par un stage à la Bibliothèque apostolique vaticane, au service de la Coordination des services informatiques, sous la direction du Dr Paola Manoni, Responsable de la coordination des services informatiques et du Dr Giuliano Giuffrida. Son travail porte sur le déploiement d'une intelligence artificielle sur les fonds manuscrits du Saint-Siège.
