Universitat de Barcelona
Grau en Informació i Documentació
Curs 2017-2018

Treball Final de Grau

Avaluació de processos de reconeixement d'entitats (NER) com a complement a interfícies de recuperació d'informació en dipòsits digitals complexos

Tutor: Josep Àngel Borràs Vela

Alumne: Gerard Vidal Santos

Agraïments:

Als meus pares
A les meves companyes
A Constança i en Josep Àngel, pels seus consells

i a tots aquells que de manera conscient, inconscient o contraproduentment
m'ensenyen maneres de veure el que m'envolta

Espero seguir aprenent de tots vosaltres

Taula de Continguts

Índex de Taules i Figures

Taules

Figures

0. Resum executiu

L’objectiu d’aquest estudi és la creació automàtica de punts d’encapçalament per mitjà de tècniques de reconeixement d’entitats (*Named-Object recognition* NER) en un conjunt de registres bibliogràfics extrets d’un agregador de tesis doctorals que poden ser relacionats directament o indirecta amb el món de la cuina per determinar la seva validesa en l'assistència al desenvolupament de robustos models de representació de coneixement en plataformes d’agregació de continguts acadèmics .

Per a tal propòsit l’estudi recopila de forma selectiva la bibliografia existent sobre experiències en l’ús d’aquest tipus de tecnologies en entorns d'àmbit bibliotecari i arxivístic, centrant-se especialment en les pautes establertes per dos articles que aborden aquesta tasca des de dos punts de vista complementaris:

- El primer, *Exploring entity recognition and disambiguation for cultural heritage collections* (van Hooland et al., 2013), proposa la avaluació de tres serveis de d'extracció d'entitats per mitjà d'una eina creada expressament per al seu funcionament en un entorn controlat.

- El segon, *Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples (Zeng, 2014)* estableix les pautes d'anàlisi de resultats basat en tres nivells (identificació- descripció-interpretació) per a l'anàlisi dels resultats que aquest tipus de tractament genera.

A traves de les pautes marcades per la bibliografia es construeix un entorn de treball per extreure i analitzar les entitats detectades per DBPedia Spotlight (el servei NER emprat per a la extracció) al conjunt de registres bibliogràfics.

Els resultats mostren que, a nivell quantitatiu, la capacitat de l’eina permet fer visible una gran quantatitat de descriptors (Noms personals o corporatius, events, enclavaments geogràfics i matèries) que permeten contextualitzar millor els registres en ser combinades amb les paraules clau ja indexades, malgrat no tenir la consistència necessària en el processos de generació d’aquestes entitats per passar el filtre de qualitat establert en la taula d’avaluació.

Aquest contratemps, però, condiciona de manera relativa la possibilitat de millorar la visibilitat dels registres en els processos de recuperació. En coordinar aquest tipus de procés tècnic amb la base semàntica que gestiona el servei d’extracció, una base de dades semàntica construïda a partir de Wikipedia que permet fer construccions lògiques a partir dels nodes que es relacionen amb la entitat extreta, les possibilitats de seguir millorant el context d’un registre a partir del nuvol de dades obertes enllaçades (Linked-Open Data, LOD) pot establir un punt permanent de contacte per seguir incrementant les opcions de filtre i descobriment en col·leccions digitals complexes i, a la vegada, permetre la revisió crítica dels materials prèviament indexats per millorar la seva experiència d’ús.

Motivacions de l’estudi

1. Les interfícies de recuperació d’informació en la era Google

"*Three parts constitute the Google Search software: a spider, a BigTable database (DB) and an interface. The first scans the Web for word presence, the second indexes and stores the information, and the third allows users to access the information. The indexing is done per word, so that each word has a quantity of resources (e.g., web pages, images or audio files) related to it.*

When a user types a certain word (or a combination of words) in the search box, Google scans its DB and returns each resource connected to that word (or combination of words) in the form of a result — a link to the site where it appears. Groys views this as a disintegration of texts into a succession of freestanding words, which turns discourses into word clouds, no longer expressing an idea, but simply comprising or not comprising a certain word (Groys 7).

Thus, avers Groys, the liberation of individual words from their grammatical structure eradicates the difference between an affirmative and a critical position, inducing the commutation of a linguistic operation (of affirmation or negation) for an extra-linguistic one (of inclusion or exclusion of words in contexts) — i.e., word curatorship (11-12)."[1]

L’impacte de la cerca simple «escriu a a caixa de text el que necessites» i els complexos algoritmes de rellevància (on es combinen models de co-ocurrència de termes amb altres categories conceptuals ponderades com la relació de determinades cadenes de text en un corpus de documents, el re-aprofitament dels historials de cerca o l’ús d’altres dades geopolítiques i socioculturals referents a l’usuari) ha condicionat de manera dràstica el comportament de cerca davant de tot sistema de recuperació d’informació digital.

Els orígens d’aquest fenomen, com exemplifica la cita anterior d’Anita Paz, ens remeten a l’adopció del patró «paraula» en l’ús d'interfícies de recuperació sense un context que permeti la valoració critica més enllà del simple còmput quantitatiu. El model, a pesar del flagrant biaix que demostra en l'anàlisi al detall d’elements sensibles en el seu modelatge (Noble, 2018), és altament efectiu en el seu conjunt i una ràpida resposta a qualsevol consulta han permès posicionar-lo com a referent en les preferències de cerca de tot usuari i establir per defecte el sistema de rellevància basat en el càlcul de probabilitats en la visualització de resultats en detriment del modelatge de sistemes d’organització del coneixement, premiant la immediatesa, el consum ràpid i les estratègies de posicionament

Aquests motius esdevenen la clau que ha facilitat la mimetització (amb diferents graus d’encert) a les interfícies de cerca i, malgrat la falta de transparència dels mecanismes que la componen, les poderoses qüestions ètiques que es generen al voltant de l’ús d’informació personal d’usuaris en el seu funcionament o la opacitat dels termes i condicions en l’intercanvi amb proveïdors d’aquestes dades, és ben habitual trobar-lo en centres bibliotecaris i arxivístics,

1 PAZ, Anita, 2013. In cerca del significato: la parola scritta nell’epoca di Google. JLIS.it [online]. 1 July 2013. No. 2. [Accessed 9 February 2018]. DOI 10/gcwfgj. Available from: https://www.jlis.it/article/view/8798

on la construcció de models de representació del coneixement basats en el punt d'accés ha estat històricament el model imperant de recuperació des de la seva creació.

La introducció d'aquest model en els sistemes de recuperació d'aquests centres s'ha contemplat favorablement en un procés d'assimilació tecnològic que han permès, entre altres bondats, la digitalització i bolcament de grans quantitats de documents a text complet per a ser consultats de manera no presencial a traves d'aquestes eines i els estudis d'usuaris corroboren que el model és altament utilitzat en la cerca sobre allò conegut (o *known item*).

Tanmateix, els mateixos estudis també revelen que els processos de cerca, de concepció cíclica i condicionats tant per la rapidesa del mitjà com el caràcter reactiu de l'usuari[2], perden efectivitat degut al gran nivell de soroll que, paradoxalment, es crea en aplicar el patró «*word*» en sistemes complexos com són les quantitats ingents de textos agregats als índexs de les bases de dades que populen aquests sistemes de recuperació, generant de retruc greus desequilibris en l'accés als recursos per la poca transparència del seu sistema de posicionament en la visualització de resultats.

Les reaccions a la poca traçabilitat del sistema de rànquing en la exploració de catàleg ha estat el punt de partida de nombrosos estudis provinents de l'àmbit de les ciències de la documentació que podem englobar en aquests tres punts:

- Textos teòrics dedicats a compilar i donar visibilitat al biaix implícit en aquest tipus d'eines (Reidsma, 2018; Kitchin, 2017; Noble, 2018)
- Casos pràctics on desenvolupen una metodologia de treball per identificar i contrarestar les seves irregularitats (Ghaphery, 2016; Clark, 2008)
- L'anivellament entre la descoberta/*browsing* i la cerca per text mitjançant la creació automàtica o semi-controlada de punts d'accés per, si no invertir la tendència, ser capaç d'oferir una alternativa viable a la recuperació per paraula-clau en aquests oceans de text.

Molts estudiosos poden concórrer en que el problema radica en la falta de preparació dels usuaris i la necessitat de programes de formació en alfabetització informacional que puguin ajudar a establir estratègies efectives de recuperació d'informació però també cal ser conscients que aquesta es només una solució parcial que no ajuda a prevenir l'augment de la complexitat de càlcul en la recuperació a text complet. Argument que juga a favor d'aquesta tercera via d'investigació i que Marcia Zeng desenvolupa de la següent manera:

2 Definició del comportament de cerca d'ususari segons Martin Hearst. Disponible a: HEARST, Marti, 2011. User interfaces for search. In: *Modern Information Retrieval* [online]. 2. New York: Addison Wesley. p. 21–55. ISBN 978-0-321-41691-9. Available from: http://grupoweb.upf.edu/mir2ed/pdf/chapter2.pdf

«Many researchers and practitioners argue that keyword searching or user-generated tags make controlled vocabularies obsolete, inefficient, and unnecessary. Yet, Gross and Taylor (2005) discovered out that over one third of records retrieved through keyword searches are those where keywords were found in subject headings. The lack of controlled vocabularies would therefore seriously affect keyword searching, the predominant way users now search for information. William Badke (2012) sees the solution in user education, particularly in the academic environment, and concludes rather pessimistically: "If we fail to advocate and if we do not restore the prominence of such vocabularies, they will disappear because of disuse and a negative cost-benefit analysis"[3]

Malgrat la bona disposició dels vocabularis controlats per al transvasament de coneixements en la comunitat científica (sovint d'arrel jeràrquica i descontextualitzada de les fonts que la generen (Yi et Mai Chan, 2009), el text de Zeng pren tot el sentit en aquelles àrees temàtiques considerades menors que sovint provenen de l'alta especialització d'un ofici gremial i d'orientació pràctica (com l'enologia, el disseny o la cuina) i que no es veuen plenament representades en aquests sistemes degut a la seva tardana incorporació al món acadèmic.

És en la seva alta especialització o el desmembrament de la seva disciplina en branques del coneixement massa allunyades entre sí on l'ús d'aquests sistemes de representació són insuficients per navegar en conjunts força heterogenis, tot i haver desenvolupat un corpus teòric que explica, organitza i dóna consistència a la seva activitat. Com a resultat és converteix en tasca francament difícil establir als nostres catàlegs una visió panòptica d'aquests subconjunts sense ometre cap resultat o silenciar subconjunts que de manera latent (o fins i tot de manera retrospectiva) poden ajudar a definir millor les seves ramificacions.

2.Objectius de l'estudi

L'objectiu d'aquest estudi és la creació automàtica de punts d'encapçalament per mitjà de tècniques de reconeixement d'entitats (*Named-Object recognition* NER), replicant el treball realitzat en aquesta tercera línia de treball, en un conjunt de registres bibliogràfics extrets d'un agregador de tesis doctorals que poden ser relacionats directament o indirecta amb el món de la cuina per determinar la seva validesa en l'assistència al desenvolupament de robustos models de representació de coneixement en plataformes d'agregació de continguts acadèmics.

Per a tal propòsit l'estudi recopilarà de forma selectiva la bibliografia existent sobre experiències en l'ús d'aquest tipus de tecnologies, centrant-se especialment en les pautes establertes per dos articles que aborden aquesta tasca des de dos punts de vista complementaris:

- El primer, *Exploring entity recognition and disambiguation for cultural heritage collections* (van Hooland et al., 2013), proposa la avaluació de tres serveis de

3ZENG, Marcia, 2014. Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples. *Knowledge Organization* [online]. 1 January 2014. P. 440–451. Available from: https://digitalcommons.kent.edu/slispubs/65

d'extracció d'entitats per mitjà d'una eina creada expressament per al seu funcionament en un entorn controlat.

- El segon, *Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples (Zeng, 2014)* estableix les pautes d'anàlisi de resultats simplificant de manera entenedora el sistema de tres nivells d'interpretació d'Erwin Panofsky[4] (identificació- descripció-interpretació) per a l'anàlisi dels resultats que aquest tipus de tractament genera.

A traves de les línies pautades, es construirà un entorn de treball que serveixi per conduir la extracció d'entitats sota el supòsit d'una cerca simple de l'estil «Test de waters» sota els paràmetres abans mencionats i realitzar un anàlisi dels resultats obtinguts sota la següent premissa:

- Avaluar els resultats de processos de reconeixement d'entitats en un entorn de cerca controlat
  - Establir quantitativament el volum i la tipologia de les entitats extretes
  - Avaluar la qualitat de les entitats extretes segons el marc teòric simplificat de Panofsky

3 Context i revisió bibliogràfica[5]

3.1. Named-Entity Recognition i les dades enllaçades

Els estudis de Zeng i van Holland tracen el seu origen en el llarg recorregut que ha demostrat l'adopció de tècniques d'indexació automàtica en processos d'indexació i classificació de documents, en especial rellevància aquells que utilitzen el processament de llenguatge natural (*Natural language processing*, NLP) com a punt de partida per a l'extracció d'informació i la creació d'etiquetes descriptives de manera automàtica o semi-automàtica. Sovint aquesta línia d'investigació es conduïda conjuntament amb altres tècniques d'arrel més computacional a mesura que es fa forta la tendència de treball sobre grans conjunts de dades, com els algoritmes de *clustering* per a l'agrupació, els arbres de decisió en classificació automàtica i, més recentment, el *deep learning* o el treball en la simulació de xarxes neurals (Sanjay et Kumbhar, 2013; Smiraglia et Cai, 2017).

D'entre les subtasques en tècniques d'extracció d'informació, el reconeixement d'entitats (NER), que persegueix la localització i classificació de cadenes de text en categories predefinides de persones, organitzacions o events, però depenent de les intencions en l'ús també quantitats, valors monetaris, percentatges, etc. (Sekine et Ranchhod, 2009) ha tingut relativa fortuna com a rutina en la creació de sistemes de vigilància competitiva i creació de

4 WIKIPEDIA CONTRIBUTORS, 2016. Erwin Panofsky. *Viquipèdia, l'enciclopèdia lliure* [online]. [Accessed 25 April 2018]. Available from: https://ca.wikipedia.org/w/index.php?title=Erwin_Panofsky&oldid=17847393

bases de dades de coneixement però ha estat en conjunció amb l'adveniment de l'anomenat *Linked Data(LD)* i *Linked Open Data (LOD)* on es posa de manifest la seva rellevància al món professional de les UI en l'impuls que pot arribar a suposar per als models clàssics de classificació i representació de coneixement i a la millora en la recuperació d'informació que, a priori, representa.

La visió de Berners-Lee sobre la web semàntica, sovint titllada d'irrealitzable, comença a prendre forma de manera lenta però (estable) si considerem que la interrelació entre models ontològics és ara prou madura per retroalimentar els models conceptuals heretats i descobrir nous punts de connexió gràcies al refinament en la implantació dels engranatges que la sustenten (específicament els llenguatge de marcatge OWL i RDF). (Sawsaa, 2014).

3.2 Experiències en NER i LOD a les unitats d'informació

La necessitat en l'adopció d'aquest tipus de tecnologies es fa especialment evident en àmbits acadèmics.

La producció acadèmica, afavorida pel creixent accés a la educació superior de classes tradicionalment excloses, ha experimentat un procés entròpic que creix exponencialment on la suma de literatura cada cop més especialitzada i l'aparició de nous punts de vista provinent de la transdisciplinarietat resultant de tots dos factors anteriors afavoreix que la complexitat del càlcul de recuperació s'incrementi degut a la ingesta constant de nou contingut. En afegir a la equació el cost en temps i dedicació de la revisió iterativa dels punts d'encapçalament, tradicionalment basat en l'anàlisi intel·lectual, el resultat esdevé inassolible per moltes UI i és comprensible -però no excusable- el declivi d'aquests sistemes de representació davant de l'allau de dades.

Alguns estudis denoten que per aquesta raó gran part del treball en l'assignació de matèries no han mirat més enllà de normatives i recomanacions de bones pràctiques a nivell locals fins que aquest tipus de tecnologies ha estat prou accessible per a ser assimilades amb èxit dins dels processos tècnics de descripció i indexació (Gracy, 2015).

Molts d'ells també prenen en consideració aquest apropament a les dades enllaçades com a potenciador d'un nou grau de flexibilitat i actualització en els models utilitzant com a punt de partida els encapçalaments de matèria de la *Library of Congress* degut a la seva base precoordinada, juntament amb l'abast i profunditat, que encara manté molts elements dels llenguatge natural que la fan especialment atractiva al processament automatitzat que és

5 El procés de recerca per documentar el context d'aquest estudi ha estat conduït per processos cíclics de cerca combinada i refinament per facetes a bases de dades especialitzes i motors de cerca (LISA, SCOPUS, Google Scholar), la recerca de conceptes fundamentals en textos seminals de la disciplina i l'anàlisis de les cadenes de citació bibliografica en aquells textos pertinents sota els seguents parametres de cerca: *Automatic analysis, Automatic indexing, Discovery services, Information retrieval i Linked Data*

capaç de generar el NER tot i la dificultat que suposa el processament de la seva estructura sintàctica (Gardner, 2012).

El més habitual, però, és la extracció a partir de grans volums de textos no tractats per aconseguir una visió global de les seves particularitats. Gracy (Gracy, 2015) sistematitza en un quadre de classificació els camps susceptibles de ser utilitzats en aquest tipus d'operacions a registres arxivístics o bibliogràfics i destaca especialment els camps de text lliure com el resum o els context dins de la serie documental com a font per a la captura de descriptors no controlats.

La bibliografia al voltant de la indexació automàtica i la redirecció semàntica també identifica errors que cal preveure en aquest tipus de sistemes. Els més habituals solen ser la captura i de-referenciació d'un enllaç a conceptes homònims, sinònims o homònims que no són pertinents al context, sobretot en les operacions de lematització, i s'accentuen en el tractament d'acrònims i inicials. A pesar d'aquests inconvenients, totes les experiències solen concloure, de manera pragmàtica, que el resultat obtingut permet un relatiu grau d'èxit en la millora de les metadades descriptives i, a la vegada, una major visibilitat dels registres al catàleg (De Wilde et Hengchen, 2017).

Però l'actualització a models de representació no és l'únic camí que es planteja en l'obertura a les dades enllaçades en un entorn web.

La tendència a mantenir vocabularis controlats d'aquestes característiques perd pistonada en l'aïllacionisme per concentrar esforços en la construcció d'estratègies de col·laboració directa o indirecta[6] (Smith-Yoshimura, 2018) davant de l'evidencia que la riquesa que proporciona el solapament de descriptors de matèria no controlats (també anomenats *Folksonomies*) poden ser un camí paral·lel igualment efectiu en la recuperació d'informació i, de retruc, en la descoberta automatitzada de paral·lelismes a altres sistemes, ja sigui de manera explicita (*author keywords*) o a partir de l'anàlisi de termes provinents dels registres bibliogràfics.

Un cop assestat l'alt grau de rellevància que poden arribar a assolir l'ús combinat de descriptors lliures i el vocabulari controlat, el repte pendent a la UI sembla ser reconduïr la tradicional perspectiva idiosincràtica en la construcció de models conceptuals cap a formes més poroses a altres punts de vista, assimilant perspectives pròpies de la teoria queer i el feminisme com a element vertebrador de discurs. (Yi et May Chai, 2009) (Drabinski, 2013) (Sandler et Bourg, 2015).

3.4 Anàlisis dels treballs de base per a l'estudi

6 Exemples d'exit són Virtual International Authority File (col·laboració directa) o Wikidata (col·laboració indirecta)

Els treballs de van Hooland i Zeng profunditzen en aquesta tècnica des d'un plantejament que divergeixen en la metodologia però que resulten complementaris en l'estudi de les seves aplicacions pràctiques.

El primer, d'arrel més pragmàtica, continua els estudis previs al voltant de tècniques automatitzades per a la reconciliació de descriptors contra vocabularis controlats (van Hooland et. Al, 2013) desenvolupant una eina d'extracció automàtica configurada per consultar serveis d'extracció d'entitats de manera remota i testejant-la amb un corpus anotat manualment per detectar els encerts i les irregularitats dels resultats obtinguts sota 4 categories preestablertes (PER-Persona o Entitat Corporativa, LOC-Emplaçament geogràfic o politic, EVE-Event i MISC-Miscel·lània on s'inclouen matèries).

L'anàlisi i la discussió del text mesura les dificultats en la correcta identificació segons paràmetres sintàctics i morfosintàctics per prendre partit a la hora d'establir la rellevància de les entitats descobertes com símptoma de la seva qualitat, més enllà de la profunditat o la complexitat del concepte extret i posa èmfasi en els principis de cost-benefici que suposa per a la exploració de material textual a les unitats d'informació.

Figure 2. Panofsky's three-layer framework and the simplified layers used by CCO.
Source: Compiled based on Panofsky, 1939, p.14-15 and CCO Chapter 4.

Object of Interpretation	**Act of Interpretation**	**Equipment for Interpretation**	**Controlling principle of Interpretation**	**Simplified layers [2]**
I-*Primary or natural* subject matter – (A) factual, (B) expressional-, constituting the world of artistic motifs	*Pre-iconographical description* (and pseudo-formal analysis).	*Practical experience* (familiar with *objects* and *events*).	History of *style* (insight into the manner in which, under varying historical conditions, *objects* and *events* were expressed by *forms*).	I-Description (refer to the generic elements depicted in or by the work).
II-*Secondary* or *conventional* subject matter, constituting the world of *images*, *stories* and *allegories*.	*Iconographical analysis* in the narrower sense of the word.	*Knowledge* of *literary sources* (familiar with specific *themes* and *concepts*).	History of *types* (insight into the manner in which, under varying historical conditions, specific *themes* or *concepts* were expressed by *objects* and *events*).	II-Identification (refer to the specific subject).
III-*Intrinsic meaning* or *content*, constituting the world of '*symbolical*' *values*.	*Iconographical interpretation* in a deeper sense (*iconographical synthesis*)	*Synthetic intuition* (familiar with the *essential tendencies* of *the human mind*), conditioned by personal psychology and '*Weltanschauung*.'	History of *cultural symptoms* or '*symbols*' in general (insight into the manner in which, under varying historical conditions, *essential tendencies of –the- human mind* were expressed by specific *themes* and *concepts*).	III – Interpretation (refer to the meaning or themes represented by the subjects and includes a conceptual analysis of what the work is about).

Figura 1 : Adaptació del Model Panofsky utilitzat per Zeng

L'article de Zeng, per la seva banda, pren com a punt de partida els estudis iconològics d'Edwin Panofsky sobre representació de conceptes en obres pictòriques per avaluar l'extracció

d'entitats amb el programari *OpenCalais*[7] en mostres poc nombroses (43-44 ítems) a partir d'una taula de tres estrats de significat (identificació- descripció-interpretació). (Figura X)

Dels resultats es destaca la gran efectivitat en detectar noms personals o corporatius i aquells conceptes representats en el text (sovint emmarcats dins el primer i segon nivell de la taula de gradació) i, com en l'estudi anterior, les grans possibilitats que planteja el seu alineament amb altres vocabularis controlats en posteriors etapes de tractament.

4.Metodologia

Considerant els elements més oportuns de la revisió bibliogràfica s'estableix la següent metodologia per avaluar processos de reconeixement d'entitats en un entorn de cerca:

- Cerca i extracció de registres bibliogràfics:
  - Exportació a entorn de treball controlat
    - Anàlisis dels elements recuperats:
      - Comportament de la cerca
      - Densitat i tipologia de descriptors/matèries
- Extracció d'entitats amb NER
  - Anàlisis quantitatiu:
    - Segmentació per categories - Anotació manual de categories
    - Comparativa amb termes indexats (volum de dades) i distribució (termes ja indexats i sequenciació temporal
  - Analisis qualitatiu:
    - Establir grau de profunditat de les entitats extretes segons quadre d'avaluació descrit per Zeng

Eines de treball seleccionades

- Eina de Scrapping web: (http://webscraper.io/)
- Eina de manipulació de dades: Openrefine (http://openrefine.org/)
- Eina d'extracció NER: (https://github.com/RubenVerborgh/Refine-NER-Extension)

El programari escollit per realitzar totes les operacions de processament de dades és *OpenRefine*, una eina de tractament i normalització de dades en accés obert que llargament citada en la bibliografia i d'ús habitual en processos de normalització i modificació de dades a gran escala en unitats d'informació.

7THOMSON REUTERS, 2018. Open Calais. *Open Calais* [online]. [Accessed 7 May 2018]. Available from: http://www.opencalais.com/

L’extracció de les dades bibliogràfiques que han de ser servir per l’estudi es durà a terme mitjançant una extensió per al navegador web *Chrome* anomenada *Webscraper.io* que permet seleccionar les etiquetes HTML d’una pàgina web i descarregar-les en un fitxer CSV

Per a la extracció de les entitats s’utilitza el servei de reconeixement d’entitats desenvolupat per Ruben Verbourgh. El servei, dissenyat per a ser utilitzat en l’entorn de treball *OpenRefine*, automatitza la crida remota per interfície de programació d'aplicacions[8] a 5 d’aquest serveis: *Alchemy-API*, *WikiMeta*, *Zemanta*, *DataTXT* (serveis propietaris) i a *DBPedia Spotlight* (gratuït).

Per l’estudi s’utilitzarà únicament aquest últim.

5. Procesament

5.1 Construcció de l’entorn de treball

El conjunt de referencies a què ha donar assistència el NER és una cerca molt genèrica realitzada a *Open Acces Thesis & Dissertasions (OATD)*[9] que emula el comportament de cerca d’una cerca simple per a intentar recuperar totes aquelles tesis que estan directament o indirecta relacionades amb el fet gastronòmic per afavorir l’aparició d’un entorn complex que generi el soroll necessari per avaluar els resultats que genera el NER a partir d’una perspectiva quantitativa, que permeti la seva comparació amb el treball d’indexació manual en entorns de gran volum i, la vegada, ser font per la extracció de subconjunts que facilitin l’avaluació qualitativa de les entitats extretes a partir de la taula d’avaluació desenvolupada per Zeng.

OATD és un agregador de tesis doctorals en accés obert d’abast mundial que permet la cerca avançada per recuperar els registres i l’ús de facetes per limitar la cerca. En recol·lectar nombroses quantitats de registres provinents d’universitats d’arreu del món sense realitzar cap tractament tècnic posteriorment i no disposa d’un llenguatge documental unitari amb què poder delimitar la cerca per camp temàtic de manera acurada.

Aquesta situació és extensible a gran part de dipòsits acadèmics d’aquesta tipologia, que es soluciona en la reducció a elements comuns de descripció bibliogràfica a tot els registres recol·lectats (sovint l’esquema bàsic de *Dublin Core*) per facilitar la ingesta. Aquesta limitació, vista com un veritable problema d'estandardització que pot provocar «situacions e inconvenients no previstos» (Tramullas, 2015). En permetre la indexació no controlada de camps en diversos idiomes a la base de dades i s’altera el bon funcionament dels motors de cerca, efecte paradigmàtic dels afers en la recuperació exposats en la introducció d’aquest treball i, per tant, un candidat idoni per a proporcionar les mostres d’estudi.

8 En anglès: *Application Programming Interface* (API)
9 OATD – Open Access Theses and Dissertations, 2018. [online]. [Accessed 8 May 2018]. Available from: https://oatd.org/

És té en consideració, també, que les tesines no han estat històricament materials bibliogràfics d'àmplia distribució, sovint dipositades a la institució d'origen i distribuïdes digitalment per agregació sota els mateixos descriptors en que va ser generades, situació que fan d'aquesta tipologia documental un cas d'estudi molt adient per al reconeixement d'entitats.

La tria del fet gastronòmic com a objecte de cerca tampoc és baladí. L'origen del fet gastronòmic com a disciplina es remunta a un art gremial i ha desenvolupat un corpus teòric propi en un espai de temps recent, tot i haver estat objecte d'estudi per a comunitats científiques molt heterogènies degut a la seva presencia ubiqua, indissoluble a facetes de la condició humana però també motiu de jocs de paraules ("*cooking the books*" per referir-se a l'alteració d'exercicis comptables, "*cooking*" fabricació de paper, entre d'altres) i rellevàncies latents provinents d'altres camps d'estudi (antropologia, etnologia però també aquelles d'on es poden establir paral·lelismes en base a metodologies, processos o resultats).

Per replicar un entorn complex que serveixi de base per generar corpus accidental(Toruella et al, 1999) com el descrit en els paràgrafs anteriors es forma una estratègia de cerca simple per matèria delimitada a l'anglès a partir de termes LCSH:

Gastronomy OR *Restaurants* OR *Cooks* OR *Cooking*

- Gastronomy: http://id.loc.gov/authorities/subjects/sh85053492.html
- *Restaurants*: http://id.loc.gov/authorities/subjects/sh85113249.html
- *Cooks*: http://id.loc.gov/authorities/subjects/sh85032173.html
- *Cooking*: http://id.loc.gov/authorities/subjects/sh2010008400.html

5.2 Descripció de l'entorn de treball[10]

La cerca retorna 470 resultats en anglès, en un rang temporal força ampli (1912-2017) i provinents majoritàriament del països anglosaxons, encara que s'observa que aproximament un terç dels registres provenen d'altres parts del món. Tots ells d'una gran varietat temàtica, que representa prou bé l'entorn complex que es pretenia afavorir en la construcció de la cerca.

Entre els resultats obtinguts es detecta, entre una diversitat considerablement ampla de tòpics, un conjunt nombrós de tesines dedicades a la restauració des de una perspectiva de negoci, un subgrup que podem englobar a partir de anàlisis etnogràfic o antropològic que prenen la gastronomia com a fil conductor, estudis sociològics o de mercat, un feix important d'estudis tècnics dedicat principalment a la preservació de productes alimentaris i a la seguretat en la seva ingesta i, finalment, grups menys nombrosos dedicats a la perspectiva literària o la nutrició o l'efecte de l'alimentació en la salut de la població.

10 El procès per descarregar els registres de OATD s'ha desenvolupat de la següent manera:
- Extracció dels registres complets recuperats del repositori amb el programari *Webscraper*
- Transformació de les dades obtingudes en un fitxer csv i ingesta del fitxer al programari Openrefine

A la vegada, també s'observen registres menys rellevants o, directament, no pertinents a la intenció de la cerca que versen sobre el re-aprofitament de materials de rebuig a la cuina (p. ex. olis), solucions tecnològiques per a millorar les condicions de vida a població amb poc recursos (p. e. forns solars i altres aplicacions d'economia domèstica) o tècniques de fabricació i tractament de paper o combustibles vegetals.

En una matriu paraula-document en les àrees on apareixen més elements[11] podem identificar una distribució bastant homogènia dels termes i l'aparició d'almenys un d'ells com a matèria en totes les combinacions de la mostra, seguint el criteri proposat. (Taula 1)

	Cooking	**Restaurants**	**Gastronomy**	**Cooks**
Tittle	87	35	5	5
Abstract	134	80	7	15
Keyword	272	172	20	14

Taula 1: Matriu de distribució de termes de cerca a registres

Analitzant en detall la matriu completa[12] queda palesa la forta relació entre els quatre conceptes de l'estratègia de cerca en el conjunt recol·lectat, però és fa més evident com la sincronia es dissipa en l'apartat de paraules clau, on *cooking* i *restaurant* acumulen el gruix de dels registres i són la via per on s'obre el camp d'estudi a altres disciplines que sumen 2055 descriptors únics molt heterogenis[13], així com diferents estils de truncament i orientacions metodològiques divergents en la selecció de termes. La seva distribució és força irregular, manté un rang entre 2 i 34 valors per registre i una mitjana de 6-7 descriptors per registre encara que el valor més habitual en la mostra[14] és 3 descriptors per registre.

Les característiques més representatives es sumaritzen en aquests 6 punts[15]:

- Termes simples o construccions precoordinades sobre l'esquema *LCSH*
- Termes en idiomes forans
- Autoritats personals, corporatives i acrònims
- Paraules clau sense estructura morfosintàctica definida (*folksonomies*)
- Valors numèrics o registres de classificació i acrònim

En calcular el nombre de paraules clau per l'any de publicació de la tesis (Figura 2) és quan comprovem que l'increment es notable a partir dels anys 2000 i considerablement alt en els textos més recent, replicant el cànon del model de comunicació científica digital i les tendències d'indexació del moment.

11 Es segueix les recomanacions de l'estudi de Gracy (Gracy, 2015) on destaca el resum, el títol i la secció de matèries com a possibles fonts per a la descoberta d'encapçalaments de matèria i autoritats

12 Annex 1. Matriu terme-document de la cerca

13 Annex 2. Relació de paraules clau ordenades de major a menor freqüència

14 Dades estadístiques globals: Valor mínim=1; Valor màxim=34; Mitjana=6.78; Moda=3; Mediana=5

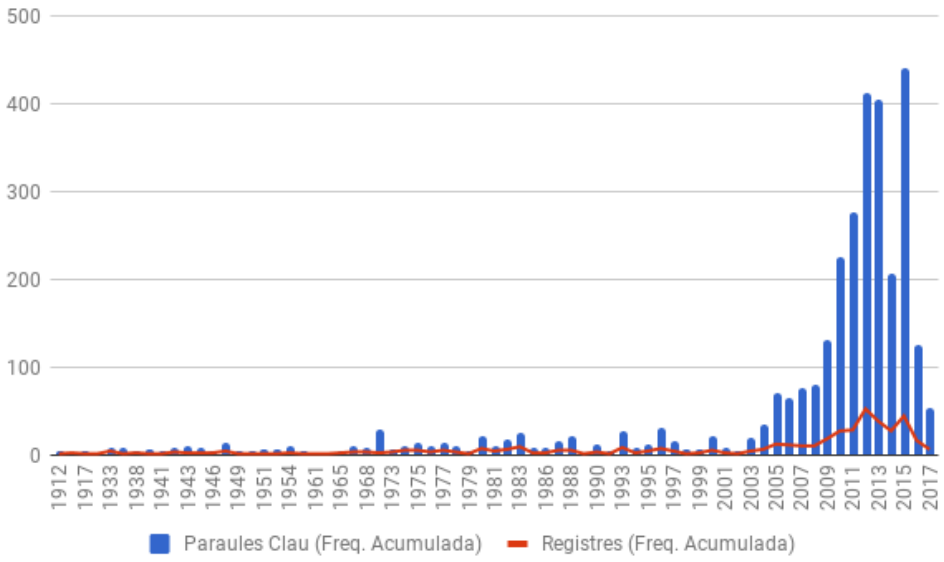


Figura 2: Distribució Paraules clau-Any

Finalment, també cal destacar que el mostreig de termes indexats té una forta relació amb el principi de Pareto, el 86% dels termes apareix només un cop en tota la mostra[16] (Annex 2)

5.3 Extració NER i descripció de les entitats extretes

Dbpedia Spotlight és una eina per a la identificació de recursos allotjats a la ontologia desenvolupada per *DBPedia* a partir de *Wikipedia* que permet enllaçar fonts d'informació no estructurada al nuvol de dades enllaçades per mitjà d'aquesta plataforma. La interfície ha estat desenvolupada per Ruben Verbough per als estudis del grup de treball en els treballs de que permet a l'usuari configurar la precisió i la confiança dels valors identificats a partir d'un quadre de configuració bàsic (Figura 3).

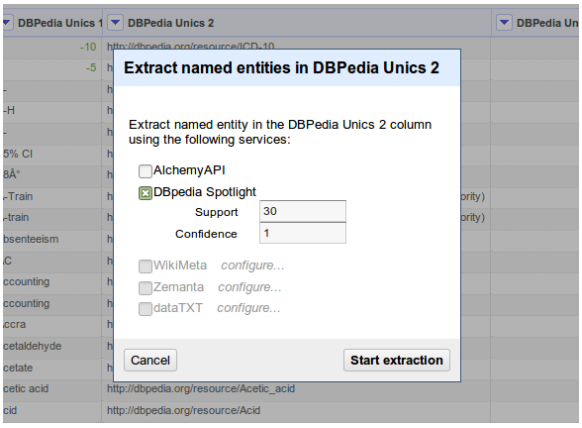


Figura 3: Panell de configuració NER-Extension

El procés de reconeixement es realitza sobre el camp *abstract* seguint les recomanacions de la revisió bibliogràfica en tres nivells de confiança, increment el valor per defecte (0.5) a 75-100%

15 Per la dificultat en la classificació al detall d'aquestes particularitats (erros tipogràfics, variacions en el truncament) no podem donar dades exactes de la seva tipologia però, a efectes pràctics podem aproximar que al voltant de un 30% dels descriptors tenen correspondència amb LCSH i assumirem aquesta dada en endavant.

16 Annex 2. Relació de paraules clau ordenades de major a menor freqüència

de confiança en detectar que no l'eina no ofereix el nivell de confiança en la detecció d'entitats per cada terme extret i recuperant 2020 descriptors únics en 332 registres[17], totes elles amb enllaç a un identificador permanent, que han estat anotats manualment sota el marc genèric en que es basen la majoria de sistemes de reconeixement (AUT; LOC, EVE; MISC)[18] tenint en compte el context que proporcionen les URIs dereferenciades a DBPedia (parant especial atenció a les etiquetes subject; type i owl:SameAs).

En comparació amb el còmput global de paraules clau s'observa una distribució superior de descriptors per registre, (Figura 4)[19]

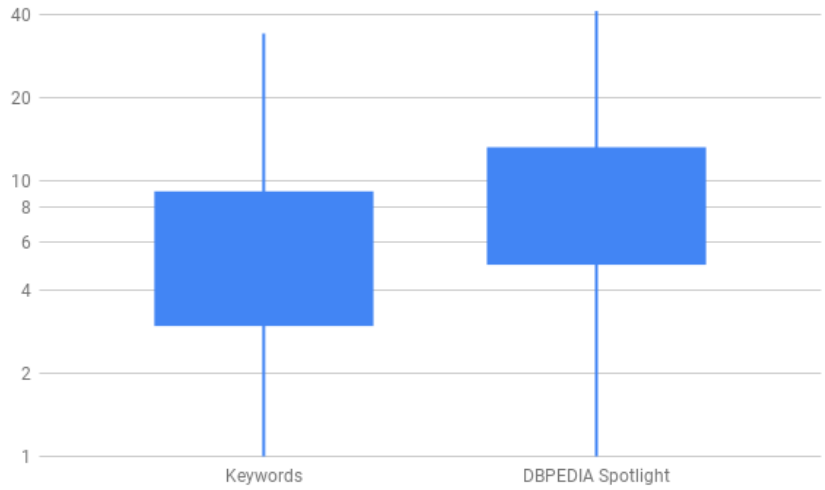


Figura 4: Distribució de freqüències Paraula Clau-DBPedia Spotlight

Els resultats (Taula 2) mostren una distribució irregular en totes tres extraccions, sent l'etiqueta dedicada a conceptes la més nombrosa (76 %), seguida per les autoritats i descriptors geogràfics (11.63% i 10.30%, respectivament) i, finalment, els events en la proporció més baixa (1.09%). En la recol·lecció a 75% de confiança es descobreixen 11 entitats no reconegudes prèviament (10 MISC, 1 LOC, 2 Errors) i en plena confiança (1%) se n'identifica 1 de nova (1 LOC, 0 Error), cap d'elles reconeguda per la versió per defecte del programa.

Taula 2: Distribució d'entitats per categories

	Freq. 0.5 (Valor per defecte)	**%**	**Freq 0.75**	**%**	**Freq. 1**	**%**
AUT	235	*11.63%*	162	*12.56%*	53	*16.01%*
EVE	22	*1.09%*	13	*1.01%*	1	*0.30%*
LOC	208	*10.30%*	194	*15.04%*	21	*6.34%*
MISC	1555	*76.98%*	921	*71.40%*	257	*77.64%*
Total	2020		1290		332	
% respecte a valor per defecte			*63.86%*		*16.44%*	

17 Annex 3. Relació d'entitats NER extretes per DBPedia Spotlight
18 AUT: Autoritats, LOC: Geografic, EVE: Events, MISC: Misce·lania(inclou matèries)
19 Dades estadístiques globals: Valor mínim=1; Valor màxim=41; Mitjana=10.08; Moda=8; Mediana=8

La taxa d'error (Taula 3) en l'assignació és bastant alta i força irregular quan la confiança en el valor retornat és de 0.5, la interfície reconeix amb un relatiu grau d'encert persones, organitzacions i accidents geogràfics i és més irregular en la categoria de conceptes, tot i que el marge baixa proporcionalment a mesura que les restriccions de confiança per al terme identificat s'incrementen.

Taula 3: Distribució d'errors per categories

	Freq. 0.5	%	Freq. 0.75	%	Freq. 1	%
AUT	124	52.77%	65	97.01%	7	100.00%
EVE	4	18.18%	0		0	
LOC	21	10.10%	0		0	
MISC	87	5.59%	2	2.99%	0	
Total	236		67		7	
% respecte a Freq.total	*11.68%*		*5.19%*		*2.11%*	

Ara bé, analitzant els resultats en detall podem confirmar que és difícil de predir el treball intern de desambiguació que *DBPedia Spotlight* realitza per decidir quins dels termes susceptibles de ser candidats són finalment triats.

5.4 Anàlisi qualitatiu

Gran part de les entitats extretes són presents al camp *abstract* i no s'aprecia cap canvi significatiu a la entitat dereferenciada de la cadena de text que ha estat identificada, més enllà dels errors i en les ocasions on els conceptes a que remeten no són exactes la entitat extreta és comparteix poc més que l'arrel del mot reconegut, donant peu a presumir que les inferències que es realitzen en el text de mostra no són suficients per establir un càlcul precís que determini la identitat del candidat i, a mesura que creix el grau de confiança en la extracció el programa adopta postures més conservadores a la hora d'establir els nexes.

Per exemple, un registre molt rellevant a la cerca obté aquests resultats:

Títol
Authorship at the fogones: Gastronomy and the Artist in Post-Transition Spain

Abstract
This dissertation analyzes the representation of food in **Spanish** novels and cookbooks from the 1980s and 90s, a period in which **Spanish cuisine** gained an unprecedented level of **international** visibility and prominence. Using both cookbooks and novels published during the period, my project examines the tension between everyday and stylized food practices in order to explore how each text engages with questions of authorship and artistic creation. The first two chapters focus on cookbooks authored by alta cocina chefs as well as by gastronomic critics who have compiled signature recipes by the chefs. I consider how these texts engage with contemporary theories of authorship and creativity, establishing a complex relationship with the modern notion of author as individual, autonomous, and unique, a kind of genius figure. In Chapter 1, I analyze the ways in which the prologue writers and compilers of two cookbooks of the early 1980s, **Carlos Delgado**'s Cien recetas magistrales (1981) and the 1982 Grandes maestros de la nueva cocina vasca present the featured chefs as unique and autonomous creators as part of a process of establishing **culinary art** as a legitimate art form.

Chapter 2 focuses on three chef-authored cookbooks of the 1990s: **Ferran Adrià**'s **El Bulli**: El sabor del Mediterráneo (1993), **Karlos Arguiñano**'s El menú de **cada día** (1992), and **Pedro Subijana'**s Menú del día (1992). I consider the problematic attempts by these professional chefs to affirm themselves as singular, creative "authors." These two

chapters identify contradictions related to the presentation of the nature of these **chefs**' "genius" and reveal unresolved tensions related to the role of the artist, of the intended reader, and also of the gastronomic critic. Despite the problematic nature of considering singular authorship within culinary creation, these texts speak to the continued legacy of the **Romantic** author and the enduring idea of a single author as originator of a unified text. Whereas these chapters on **cookbooks** consider the way in which food practices are "written" and thereby offered up for aesthetic consideration, the final two chapters consider how aesthetic objects—in this case novels—utilize food, serving both realistic and metaphorical functions, in order to contemplate what it means to be an artist and author as well as the role of creativity within the everyday.

In Chapter 3, I analyze the function of food in **Manuel Vázquez Montalbán**'s El pianista as a complex and ambiguous depository of **memory**. In this novel, culinary references prompt an exploration of how the negotiation of everyday practices in the present reveals modes of engaging with the past as well as the role and authority of the artist in contemporary society. Chapter 4 examines how the attention paid to quotidian food practices in **Almudena Grandes**' **Malena es un nombre de tango** facilitates a breaking down of **binary** oppositions, accompanied by an affirmation of creative authorship by the main character Malena. Such explorations of the meaning of everyday food practices...

Paraules clau
Literature; Cookbooks; Spain; Post-Transition; authorship; cuisine; culinary arts; Gastronomy; food practices; the artist; the everyday

DBPedia Spotlight (0.5)	DBPedia Spotlight (0.75)	DBPedia Spotlight (1)
• Spanish – Spain(LOC) • Spanish cuisine (MISC) • international - International law (MISC) (ERROR) • Carlos Delgado (AUT) (ERROR, baseball player) • culinary art – Culinay art (MISC) • Ferran Adriá (AUT) • El Bulli (AUT) • cada día – Papito_(album) (AUT) (ERROR) • Pedro – Pedro II of Brazil (AUT) (ERROR) • Romantic – Romanticism (MISC) • Manuel Váquez Montalbán (AUT) • memory (MISC) • tango – Tango music (MISC) • binary – Binary_number (MISC)	• Spanish cuisine • international • Carlos Delgado • culinary art • Ferran Adriá • El Bulli • Manuel Vázquez Montalbán • memory • tango	• Spanish cuisine • Carlos Delgado • El Bulli

DBPedia Spotlight situa bé el descriptor geogràfic trencant l'arrel de la paraula (*Spanish- Spain*) i identifica correctament 3 de les 7 autoritats possibles, descartant els titols que apareixen al text (*Ferran Adriá, El Bulli, Manuel Vázquex Montalbán*) però errròniament assigna 2 possibles candidats a persones sense cap relació amb la context de la tesis i descarta Karlos Arguiñano i Almudena Grandes en els seu valor per defecte.

En els conceptes seleccionats trobem un resultats intermitents, entre rellevant (*Spanish cuisine, culinary arts*) irellevants (*Romantic, memory, tango, binary*), i erronis (*International - international law, cada dia - Papito(album))* mentre es troben a faltar alguns conceptes relacionats amb les paraules clau indexades que són presents al text (*cookbooks, creativity, chefs entre d'altres*).

Incrementant el nivell de confiança els termes van desapareixent, mantenint un cert equilibri en un 75% de confiança i perdent alguns termes rellevants en el 100%

Aquest cas és força representatiu dels resultats obtinguts, que és poden sintetitzar de la següent manera:

- Forta presencia de descriptors poc rellevants a la cerca *(energy, memory...)*
- Termes específics o molt especialitzats són reconeguts satisfactòriament
- Abreviacions susceptibles de ser mal interpretades o reconegudes com acrònims
- Obres artístiques i autoritats assignades per error per la homonímia-similitud de la cadena de text
- Multilingüisme i paraules foranes retornen errors en la mostra però són identificades correctament si són manlleus acceptats (*Longue Duree)*

El model Panofsky que s'havia plantejat en la metodologia com a mètode per desentrellar aquesta esquiva «qualitat» en els termes extrets no es considera aplicable en aquest estudi degut a les característiques detectades durant el procés de categorització. Recordem els tres nivells de profunditat que estableix l'estudi per determinar el grau de profunditat dels descriptors:

- Nivell 1: Identificació de valors, subjectes i conceptes subjacents a la temàtica de la obra
- Nivell 2: Identificació de conceptes que representen la temàtica de la obra
- Nivell 3: Identificació de conceptes que identifiquen les motivacions o el significat de la obra

Analitzant els resultats en detall podem confirmar que és difícil de predir el treball intern de desambiguació que *DBPedia Spotlight* realitza per decidir quins dels termes susceptibles de ser candidats són finalment triats. Totes (o quasi totes) les entitats extretes han de ser categoritzades en aquest primer estadi de representació conceptual. Les inferències observades devenen d'un treball amb més fortuna que raciocini al darrere, i fins i tot es podria plantejar un nivell 0 en moltes de les entitats extretes per la irrellevància que aporten al contingut per si soles.

El segon nivell implica un treball a partir d'inferències que es podrien donar gràcies a la l'estructura de relacions que permet la ontologia de *DBPedia* però que la infraestructura tècnica de la eina no és capaç d'assolir per si sola, només en aquest cas es podria assolir la profunditat que el varem demana.

Finalment, el tercer nivell implica una conceptualització que mai arriba a assolir-se degut a les pròpies característiques de la mostra, en el cas que s'identifica les motivacions de l'autor en plantejar l'estudi el resultat només és dóna per que el text el fa explicit.

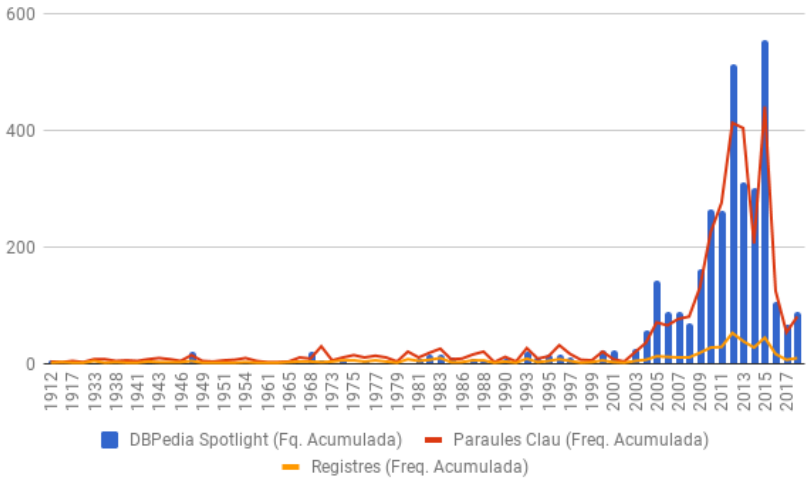
6. Conclusions i futures actuacions

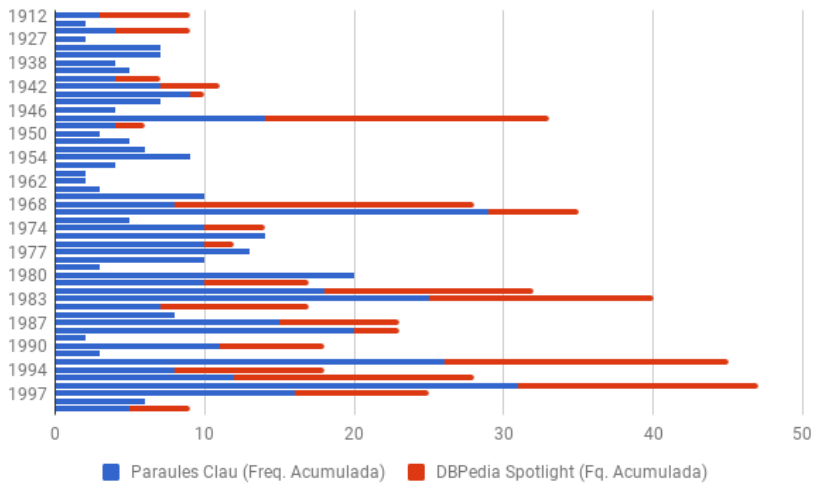
Els resultats obtinguts en aquest l'estudi aconsellen un cert grau de prudència en l'establiment d'entitats generades per aquest tipus de sistema en processos en repositoris destinats a l'agregació de continguts de múltiples fonts o altres entorns complexos.

És destacable, almenys en el pragmatisme que sustenta la generació automàtica de punts d'encapçalament, com és planteja en termes globals la distribució de punts d'encapçalament un cop afegit aquesta nova capa i les avantatges que suposa en termes cost-benefici per a la unitat d'informació a l'hora d'establir un mapa conceptual dels elements més rellevants que doni suport als descriptors originals.

En observar la nova distribució de freqüències al conjunt de resultats recuperats (figura 5), l'increment de punts d'encapçalament susceptibles de ser aprofitats per triar o desestimar un resultat augmenta considerablement, afavorint la visió panoràmica de conceptes que es veuen involucrats en la crida inicial i és especialment destacable el suport que és rep en aquells registres amb pocs descriptors (Figura 6)

Figures 5-6: Distribució DBPedia Spotlight-Any; Freq. acumulada Paraula clau-DBPedia Spotlight





En aquest supòsit, l'usuari és capaç de corregir més ràpidament la seva estratègia de cerca amb més coneixement de causa, i identificar descriptors rellevants que poden donar més visibilitat al registre envers al conjunt, descobrir rellevàncies poc evidents a necessitats d'informació molt diverses, o no contemplades en el procés d'indexació pel personal tècnic.

Ara bé, l'anàlisi qualitatiu de les dades demostra els resultats no haurien de ser considerats definitius degut a que l'amplia diversitat de punts de vista que proporcionen tant a nivell macro com en ser contextualitzades es veu inevitablement compromesa per l'alt grau de simplificació que moltes d'elles aporten a el registre d'on s'extreuen, deixant de banda matisos complexos o descartant descriptors pertinents a la descripció del text.

La taula d'avaluació qualitativa, plantejada per un sistema diferent d'extracció NER, permet establir una línia vermella en la que separar els encapçalaments que són pertinents i els que són irrellevants però no establir la legitimitat dels que tenen connotacions més enllà de la mera presencia contextual. Els mateixos autors sintetitzen en la discussió dels resultats la impossibilitat de consensuar un marc neutre on testejar el nivell de qualitat de les noves entitats generades sense deslligar cada descriptor del context on es genera o la utilitat que ha de perseguir segons l'objectiu amb que ha estat creada.

Gran part dels motius en la poca traça del programa es troben en la deguda contextualització de la font d'origen d'on procedeixen les dades. *DBPedia Spotlight* utilitza una ontologia basada en els recursos disponibles a Wikipedia per comprovar l'adequació de les paraules susceptibles de ser reconegudes i aquesta base de coneixement és un reflex de moltes de particularitats de Wikipedia com a font d'informació, un recurs en construcció permanent on certes àrees estan més desenvolupades que altres i que no té la consistència d'una ontologia creada expressament, tot i que formuli moltes de les relacions semàntiques habituals en aquests sistemes entre les entitats descrites (Mendelyan et al, 2009).

Els recents desenvolupaments tècnics implementats a *Wikipedia* permeten considerar el treball de reconeixement d'entitats com un pas intermedi cap a un nou model d'explotació de metadades descriptives en entorns complexos com OATD o grans plataformes de continguts en actuar com un port(«*hub*») que permet relacionar les entitats identificades amb altres vocabularis, comparar les relacions que s'estableixen entre elles, teixint als registres aquelles que són pertinents i permetent, a la vegada, l'avaluació crítica dels encapçalaments que sustenten el model conceptual en que es basa.

Reconstruint la idea interior amb un dels descriptors del conjunt recuperat per *DBPedia* que no era present en el bloc de paraules clau indexades (*2008 financial crisis*[20]) es possible construir relacions jeràrquiques, la traducció del concepte a 47 idiomes, sumar els termes equivalents

20 Entitat dereferenciada:Great Recession (http://dbpedia.org/resource/Great_Recession)

en les seves respectives llengües, identificar els conceptes relacionats i establir la seva equivalència en altres llenguatges controlats (Figura 7)

Figura 7. Exemples d'extensió LOD

Aquest petit exercici final permet constatar les possibilitats que s'obren en considerar l'obertura al nuvol per plantejar noves vies d'estudi que permetin avançar en la millora i actualització dels sistemes de recuperació d'informació en entorns complexos sota l'eix de la l'exploració dels punt d'encapçalament que apuntalen les relacions entre els registres que en formen part, permetent a l'usuari la navegació i la descoberta de possibles respostes a les seves necessitats de cerca a partir de punts de vista més plurals i no estacants en la idiosincràtica visió heretada de la institució que la genera.

7. Bibliografia

CLARK, Jason A., 2008. Making Patron Data Work Harder: User Search Terms as Access Points? *Code4Lib Journal*. 1 June 2008. No. 3, p. 78. http://journal.code4lib.org/articles/11355

DE WILDE, Max and HENGCHEN, Simon, 2017. Semantic Enrichment of a Multilingual Archive with Linked Open Data. [online]. 2017. [Accessed 16 April 2018]. Available from: https://helda.helsinki.fi/handle/10138/233900

DRABINSKI, Emily, 2013. Queering the Catalog: Queer Theory and the Politics of Correction. *Brooklyn Library Faculty Publications* [online]. 1 January 2013. Available from: https://digitalcommons.liu.edu/brooklyn_libfacpubs/9

GARDNER, Sue Ann, 2012. Cresting toward the Sea Change. *Library Resources & Technical Services* [online]. 1 April 2012. Vol. 56, no. 2, p. 64–79. [Accessed 22 April 2018]. DOI 10/gdcck5. Available from: https://journals.ala.org/lrts/article/view/5565

GHAPHERY, Jimmy, OWENS, Emily, COGHILL, Donna, GARIEPY, Laura, HODGE, Megan, MCNULTY, Thomas and WHITE[1, Erin, 2016. Building Bridges with Logs: Collaborative Conversations about Discovery across Library Departments. *The Code4Lib Journal* [online]. 25 April 2016. No. 32. [Accessed 2 February 2018]. Available from: http://journal.code4lib.org/articles/11355

GRACY, Karen F., 2015. Archival description and linked data: a preliminary study of opportunities and implementation challenges. *Archival Science* [online]. September 2015. Vol. 15, no. 3, p. 239–294. [Accessed 22 April 2018].DOI 10.1007/s10502-014-9216-2.Available from: http://link.springer.com/10.1007/s10502-014-9216-2

HEARST, Marti, 2011. User interfaces for search. In: *Modern Information Retrieval* [online]. 2. New York: Addison Wesley. p. 21–55. ISBN 978-0-321-41691-9. Available from: http://grupoweb.upf.edu/mir2ed/pdf/chapter2.pdf

KITCHIN, Rob, 2017. Thinking critically about and researching algorithms. *Information, Communication & Society*. 2 January 2017. Vol. 20, no. 1, p. 14–29. DOI 10/gc3hsj.

MEDELYAN, Olena, MILNE, David, LEGG, Catherine and WITTEN, Ian H., 2009. Mining meaning from Wikipedia. *International Journal of Human-Computer Studies* [online]. September 2009. Vol. 67, no. 9, p. 716–754. [Accessed 18 February 2018]. DOI 10/bfnk2v. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1071581909000561

NADEAU, David and SEKINE, Satoshi, 2009. A survey of named entity recognition and classification. In: *Named Entities: Recognition, classification and use* [online]. Amsterdam ; Philadelphia: John Benjamins Pub. Company. p. 3–28. 19. [Accessed 15 May 2018]. ISBN 978-90-272-8922-3. Available from: https://nlp.cs.nyu.edu/sekine/papers/li07.pdf

NOBLE, Safiya Umoja, 2018. *Algorithms of oppression: how search engines reinforce racism*. ISBN 978-1-4798-3724-3.

REIDSMA, Matthew, [no date]. Algorithmic Bias in Library Discovery Systems. [online]. [Accessed 17 February 2018]. Available from: https://matthew.reidsrow.com/articles/173

SANJAY, Desale and KUMBHAR, Rajendra, 2013. Research on Automatic Classification of Documents in Library Environment: A Literature Review. Knowledge Organization. 1 January 2013. Vol. 40, p. 295. Avaliable from: https://www.researchgate.net/publication/268505273_Research_on_Automatic_Classification_of_Documents_in_Library_Environment_A_Literature_Review

SADLER, Bess and BOURG, Chris, 2015. Feminism and the Future of Library Discovery. *The Code4Lib Journal* [online]. 15 April 2015. No. 28. [Accessed 1 February 2018]. Available from: http://journal.code4lib.org/articles/10425

SAWSAA, Ahlam and JOAN, Lu, 2014. Using Natural Language Programming (NLP)Technology to Model Domain Ontology OTO by Extracting Occupational Therapy Concepts - University of Huddersfield Repository. *Knowledge Organization* [online]. 2014. Vol. 46, no. 6, p. 452–464. [Accessed 9 May 2018]. Available from: http://eprints.hud.ac.uk/id/eprint/24354/

SMIRAGLIA, Richard P. and CAI, Xin, 2017. Tracking the Evolution of Clustering, Machine Learning, Automatic Indexing and Automatic Classification in Knowledge Organization. *Knowledge Organization : KO; Wuerzburg* [online]. 2017. Vol. 44, no. 3. [Accessed 10 December 2017]. Available from: https://search.proquest.com/lisa/docview/1927133404/CB9C71BDF158428EPQ/1

PAZ, Anita, 2013. In cerca del significato: la parola scritta nell'epoca di Google. [online]. 1 July 2013. No. 2. [Accessed 9 February 2018]. DOI 10/gcwfgj. Available from: https://www.jlis.it/article/view/8798

SMITH-YOSHIMURA, Karen, 2018. Are distributed models for vocabulary maintenance viable? – Hanging Together. [online]. 2018. [Accessed 16 April 2018]. Available from: http://hangingtogether.org/?p=6672

TRAMULLAS, Jesús, [2015]. Hispana: una revisión crítica | blok de bid. *Blok de BiD* [online]. [Accessed 29 April 2018]. Available from: http://www.ub.edu/blokdebid/ca/node/593

VÁLLEZ, Mari, PEDRAZA-JIMÉNEZ, Rafael, CODINA, Lluís, BLANCO, Saúl and ROVIRA, Cristòfol, 2015. Updating controlled vocabularies by analysing query logs. *Online Information Review* [online]. 9 November 2015. Vol. 39, no. 7, p. 870–884. [Accessed 30 April 2018]. DOI 10/f78wgq. Available from: http://www.emeraldinsight.com/doi/10.1108/OIR-06-2015-0180

VAN HOOLAND, Seth, DE WILDE, Max, VERBORGH, Ruben, STEINER, Thomas and VAN DE WALLE, Rik, 2013. Exploring entity recognition and disambiguation for cultural heritage collections. *DIGITAL SCHOLARSHIP IN THE HUMANITIES*. 1 November 2013. Vol. 30, no. 2, p. 262–279. DOI 10.1093/llc/fqt067

YI, Kwan and MAI CHAN, Lois, 2009. Linking folksonomy to Library of Congress subject headings: an exploratory study. *Journal of Documentation* [online]. 16 October 2009. Vol. 65, no. 6, p. 872–900. [Accessed 9 November 2017]. DOI 10.1108/00220410910998906. Available from: http://www.emeraldinsight.com/doi/10.1108/00220410910998906

ZENG, Marcia, 2014. Using a Semantic Analysis Tool to Generate Subject Access Points: A Study Using Panofsky's Theory and Two Research Samples. *Knowledge Organization*. 1 January 2014. P. 440–451

8. Annexos

Annex 1. Matriu terme-document de la cerca

	Cooking	Restaurants	Gastronomy	Cooks
Tittle	87	35	5	5
Abstract	134	80	7	15
Keyword	272	172	20	14

Combinació dels termes	Frequencia	%
Cooking (keywords)	104	22.13%
Restaurants (keywords)	91	19.36%
Cooking (abstract) + Cooking (Keywords)	57	4.68%
Cooking (title) + Cooking (abstract) + Cooking (Keywords)	52	8.72%
Restaurants (abstract) + Restaurants (keywords)	41	8.72%
Cooking (title) + Cooking (Keywords)	27	1.91%
Restaurants (title) + Restaurants (abstract) + Restaurants (keywords)	22	4.68%
Restaurants (title) + Restaurants (keywords)	9	1.91%
Gastronomy (keywords)	6	1.28%
Cooks (keywords)	4	0.85%
Gastronomy (abstract) + Gastronomy (keywords)	4	0.85%
Cooking (abstract) + Cooking (Keywords) + Cooks (abstract)	4	0.85%
Gastronomy (title) + Gastronomy (keywords)	3	0.64%
Cooking (Keywords) + Restaurants (abstract)	3	0.64%
Cooks (abstract) + Cooks (keywords)	2	0.43%
Gastronomy (title) + Gastrnomy (abstract) + Gastronomy (keywords)	2	0.43%
Restaurants (abstract) + Cooks (keywords)	2	0.43%
Cooking (abstract) + Gastronomy (keywords)	2	0.43%
Cooking (abstract) + Restaurants (title) + Restaurants (abstract) + Restaurants (keywords)	2	0.43%
Cooking (abstract) + Cooking (Keywords) + Restaurants (abstract)	2	0.43%
Restaurants (abstract) + Gastronomy (abstract) + Gastronomy (keywords)	1	0.21%
Restaurants (title) + Restaurants (abstract) + Restaurants (keywords) + Gastronomy (keywords)	1	0.21%
Cooking (Keywords) + Cooks (abstract)	1	0.21%

Cooking (Keywords) + Cooks (title)	1	0.21%
Cooking (Keywords) + Cooks (title)	1	0.21%
Cooking (Keywords) + Gastronomy (keywords)	1	0.21%
Cooking (abstract) + Cooks (title) + Cooks (abstract) + Cooks (keywords)	1	0.21%
Cooking (abstract) + Restaurants (keywords)	1	0.21%
Cooking (abstract) + Restaurants (abstract) + Restaurants (keywords)	1	0.21%
Cooking (abstract) + Cooking (Keywords) + Cooks (keywords)	1	0.21%
Cooking (abstract) + Cooking (Keywords) + Cooks (title) + Cooks (abstract)	1	0.21%
Cooking (abstract) + Cooking (Keywords) + Cooks (title) + Cooks (abstract) + Cooks (keywords)	1	0.21%
Cooking (abstract) + Cooking (Keywords) + Restaurants (keywords)	1	0.21%
Cooking (title) + Cooking (abstract) + Restaurants (abstract) + Restaurants (keywords) + Cooks (abstract)	1	0.21%
Cooking (title) + Cooking (abstract) + Cooking (Keywords) + Cooks (keywords)	1	0.21%
Cooking (title) + Cooking (abstract) + Cooking (Keywords) + Cooks (abstract)	1	0.21%
Cooking (title) + Cooking (abstract) + Cooking (Keywords) + Restaurants (abstract)	1	0.21%
Cooking (title) + Cooking (abstract) + Cooking (Keywords) + Restaurants (abstract) + Cooks (abstract)	1	0.21%

Annex 2.Freqüencia de paraules clau en registres capturats:

Disponible a:

https://docs.google.com/spreadsheets/d/14XhYth05bwyKQrEfu5opphZdClY9rH7zlURMqZjoejs/edit?usp=sharing

Annex 3. Relació d’entitats NER extretes per DBPedia Spotlight

Disponible a:

https://docs.google.com/spreadsheets/d/1nRjPnP-kAY0VAPqZAROxdYFYzzYYeMMlSipCSED5muY/edit?usp=sharing