



Qualinca

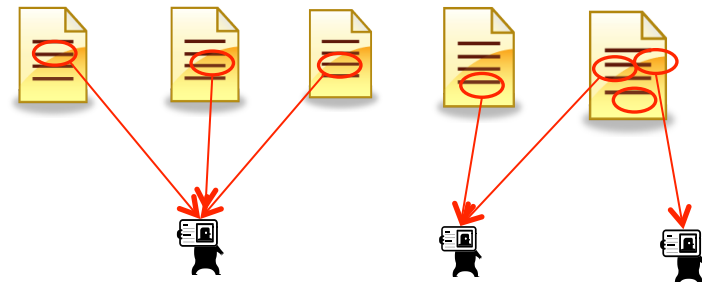
Qualité et interopérabilité
de grands catalogues
documentaires

Contexte

- Catalogues documentaires
 - Notice documentaire : des métadonnées décrivant un document en utilisant des **références d'autorité** pour désigner les entités nommées utiles à cette description
 - Notice d'autorité : des métadonnées décrivant une entité nommée à laquelle est associée une **référence d'autorité**
- Objectifs
 - S'assurer de la correction et de la complétude de l'utilisation des références d'autorité dans les notices documentaires
 - Doter les autorités d'un niveau de description suffisant pour permettre l'identification de l'entité nommée qu'elle représente.
 - Fournir des méthodes d'aide à la résolution de divers problèmes de résolution d'entité dans le contexte des catalogues documentaires.
- Enjeux
 - Disposer d'un catalogue de qualité garantissant une exploitation correcte de son contenu quels qu'en soit les usages (indexation, recherche documentaire, autre exploitation sémantique dédiée, exposition dans le web de données...)

Numéro de notice : 068246730
 Titre : Une Méthode de développement mise à la portée de tous les paysans : le ferment relationnel du développement / P. Boisseau
 Alphabet du titre : latin
 Auteur(s) : **Boisseau, Pierre, 058757988** Auteur
 Date(s) : [ca 1976]
 Langue(s) : français
 Pays : France
 Editeur(s) : [S.L.] : Revue Purpan, [ca 1976]
 Description : P. 55-75 ; 30 cm
 Notes : Notice réd. d'après la couv. - Extr. de : "Revue Purpan", n° 98, p. 55-75
 Origine de la notice : PRITEC
 Lien(s) externe(s)
 Worldcat : 492112315

058757988
 Forme retenue
Boisseau, Pierre
 Information
Langue d'expression : Français
Nationalité : France
Date de naissance : 19
Sexe : masculin
 Notes
 Docteur en droit public, Maître de conférences à l'Université de Tours (2000)
 Source
 La Commune de Paris de 1871 à l'épreuve du droit constitutionnel /
 Pierre Boisseau ; préf. Jean Rossetto ; avant-propos Jean-Pierre Massias, 2000
 Informations sur la notice
 Dernière modification : 06-10-2009 à 04 h 28



Partenariat

- Partenaires en informatique
 - LIRMM (GraphiK) : langages de représentation de connaissances et mécanismes de raisonnement adaptés au développement de systèmes à base de connaissances.
 - LRI (IASI) : intégration sémantique de données hétérogènes en particulier dans le cadre du web sémantique.
 - LIG (HADAS) : gestion de grandes bases de données dans un distribué et hétérogènes en particulier dans le cadre du web sémantique.
- Partenaires en science documentaire
 - ABES : construction et exploitation de métadonnées bibliographiques sur les ressources documentaires de l'enseignement supérieur
 - INA (CAP) : construction et exploitation de métadonnées audiovisuelles sur les contenus des archives nationales de l'audiovisuel

Problèmes abordés

- Formalisation du contenu des catalogues afin de définir
 - une notion de « qualité » d'un catalogue documentaire
 - différents scénarios d'usages des catalogues mettant en jeu cette notion de qualité (vérification, insertion, fusion...)
- Développement de méthodes et outils adaptés aux scénarios d'usage envisagés
 - résolution d'entités : *deux références représentent-elles la même entité, des entités différentes ?*
 - enrichissement/fusion de descriptions d'entités
 - découverte de clés
- Mise en œuvre des scénarios proposés sur les données de l'ABES et l'INA
 - transformation des catalogues dans la formalisation choisie
 - paramétrage des méthodes proposées => captation de l'expertise des catalogueurs
 - expérimentation/évaluation des méthodes
 - spécification/réalisation de démonstrateurs métiers

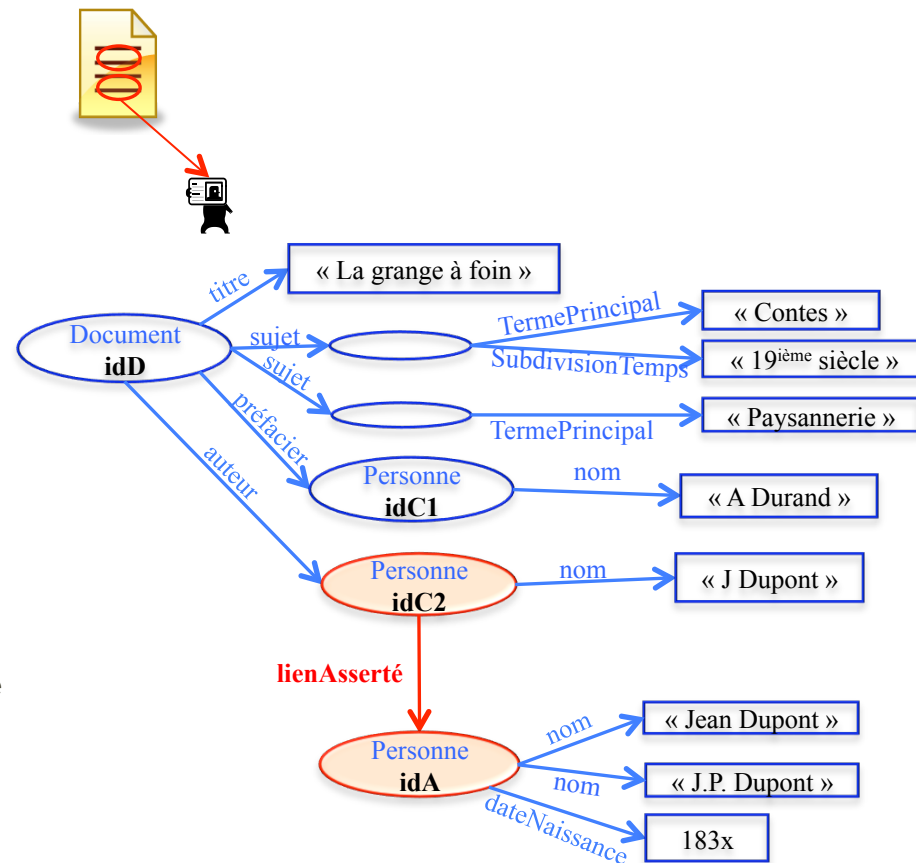
Plan

- Cadre de travail
 - Cadre formel
 - Notion de qualité
 - Les problématiques ABES et INA
- Méthodes proposées
 - Découverte de clés
 - Réconciliation de références par enrichissement à partir du LOD
 - Vérification des liens par partitionnement
- Résultats et Perspectives

Cadre de travail

Formalisation d'un catalogue en base de connaissances

- Choix d'une **ontologie** pour exprimer les métadonnées
- Transformation des notices en une base de **faits** utilisant ce vocabulaire
- Prise en compte des « erreurs potentielles » de liage
 - Un identifiant par notice d'autorité : une **référence d'autorité** d'entité nommée
 - Un identifiant par mention d'entité nommée dans une notice documentaire : une **référence contextuelle** d'entité nommée
 - Une relation « **lienAsserté** » entre référence contextuelle et référence d'autorité



Démarche cœur de projet

- A partir de l'ontologie, on élabore DES **RÈGLES D'IDENTIFICATION/ DIFFÉRENTIATION** de 2 références

$$\text{Prop}[x], \text{Prop}'[y], \text{Critères}[x, y] \rightarrow x=y \quad (\text{ou } x \neq y)$$
 - propriétés sur les références à comparer

Ex : une personne ayant un rôle scientifique dans un document contemporain
 - critères de comparaison de deux références selon une dimension sémantique particulière

Ex : Noms, Thématique, Période
- Les **propriétés** indiquent les références pertinentes pour une règle
- Les **critères** correspondent à des prédicats calculés retournant un indice de **similarité/dissimilarité** selon une dimension

Ex : MêmeNom, NomsCompatibles, NomsEloignés, NomsIncompatibles
- Les **dimensions** sont incarnées par des **attributs** des références
 - extraits par des **requêtes** sur la base

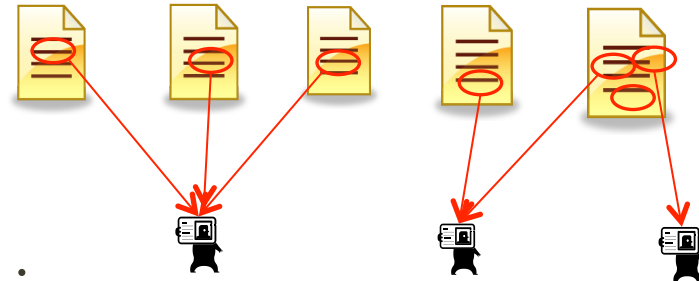
Ex : $Q(x, t) \text{ :- } \text{personne}(x), \text{rôleScientifique}(x, y), \text{sujet}(y, z), \text{termePrincipal}(z, t)$
 - inférés à partir des **liens sûrs** par **aggrégation** d'attributs d'autres références

Ex : périodeActivité d'une autorité à partir des datePublication des contextuelles

Qualité d'un catalogue

Qualité

- Hypothèse fondamentale :
 - Les seules erreurs possibles concernent les liens entre réf contextuelles et réf d'autorité
- Propriétés désirées
 - Complétude du liage : toute réf contextuelle c est liée à une réf d'autorité $A(c)$
 - Correction du liage : c et $A(c)$ représentent bien la même entité
 - Unicité des autorités : deux réf d'autorité différentes $a1$ et $a2$ représentent des entités différentes
 - Complétude de la base : les connaissances associées aux réf contextuelles et d'autorité sont suffisantes pour garantir les propriétés ci-dessus



Formalisation

- On suppose que l'on dispose de prédicats **id** et **di** sur la base **K** tels que
 - si $S \vdash \text{id}(r, r')$ alors les références **r** et **r'** représentent la même entité (avec une « bonne » confiance)
 - si $S \vdash \text{di}(r, r')$ alors les références **r** et **r'** représentent des entités différentes (avec une « bonne » confiance)
- On utilise ces prédicats pour établir ou approximer les propriétés précédentes en s'appuyant sur des :
 - notions classiques (Clé, GRE...)
 - nouvelles notions (UNP, bien fondé...)
- On étudie l'instanciation de ces notions dans le fragment logique Datalog+/-

La problématique INA

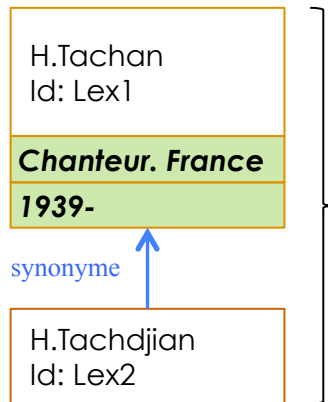
Données, Terminologies

- 11 Millions de notices de contenus
- Lexiques de 800 000 « termes » (label + notes textuelles)
- Constitution d'une ontologie OWL pour représenter l'ensemble des données de l'Ina et des concepts extraits tels que interprétation, œuvre, occupation ...
- Constitution d'une terminologie de métiers pour guider l'extraction
 - sources: WikiCatégories, Yago2, Ina, code ROME
- Mise à niveau de bases du LOD pour préparer l'étape de liage
 - Traduction en RDFS/OWL de tables choisies de MusicBrainz (œuvres musicales)
 - Complétion de Dbpédia.fr au niveau des métiers (occupation)

Construction du corpus INA

- Objectif: Structurer et enrichir ce corpus sous un formalisme facilitant l'étape de liage.
- Création d'autorités enrichies à partir des lexiques.
- Création de références contextuelles enrichies à partir des notices de contenus.

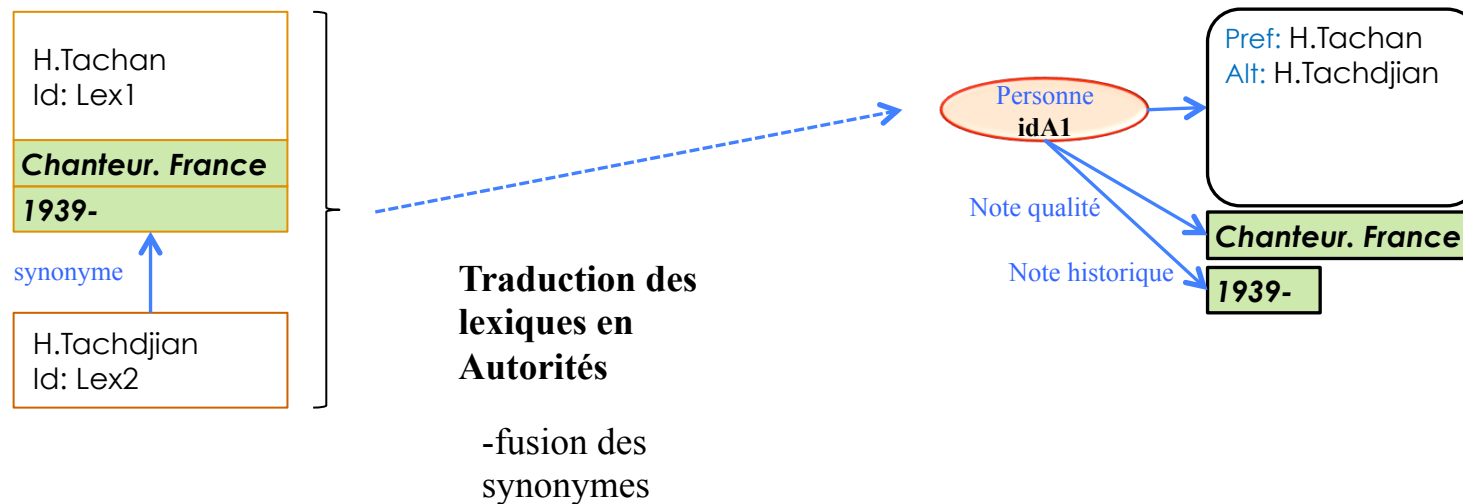
Du terme lexical vers l'Autorité



**Traduction des
lexiques en
Autorités**

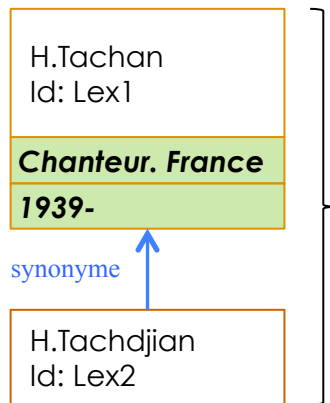
J.Lesage Id:Lex3
« homonymes: 1-Homme Politique. Canada 2-Auteur musical. Canada

Du terme lexical vers l'Autorité



J.Lesage
Id:Lex3
« homonymes:
1-Homme Politique. Canada
2-Auteur musical. Canada

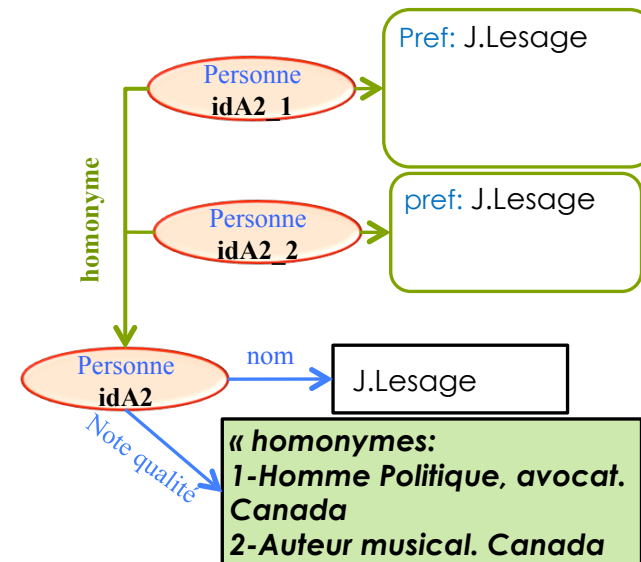
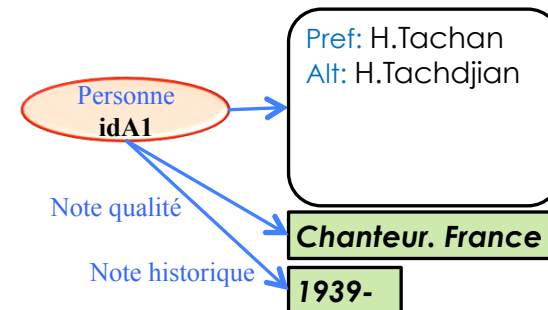
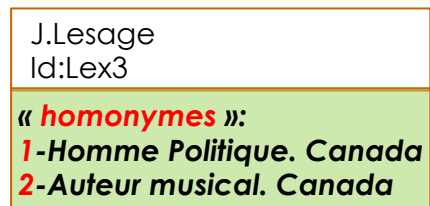
Du terme lexical vers l'Autorité



Traduction des lexiques en Autorités

-fusion des synonymes

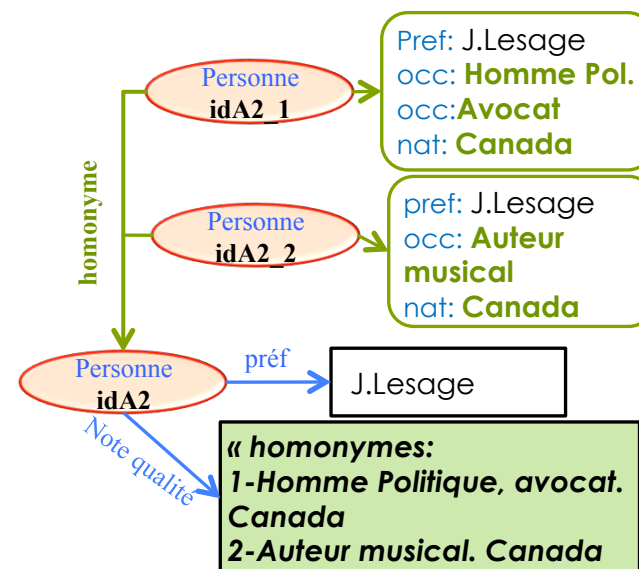
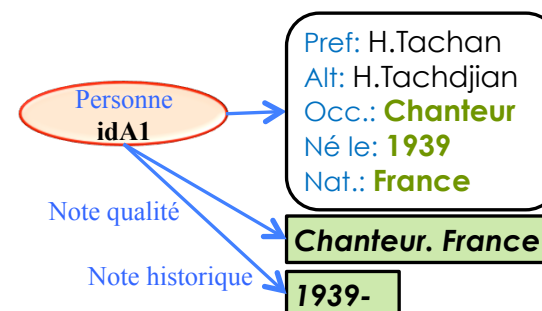
- séparation des homonymes



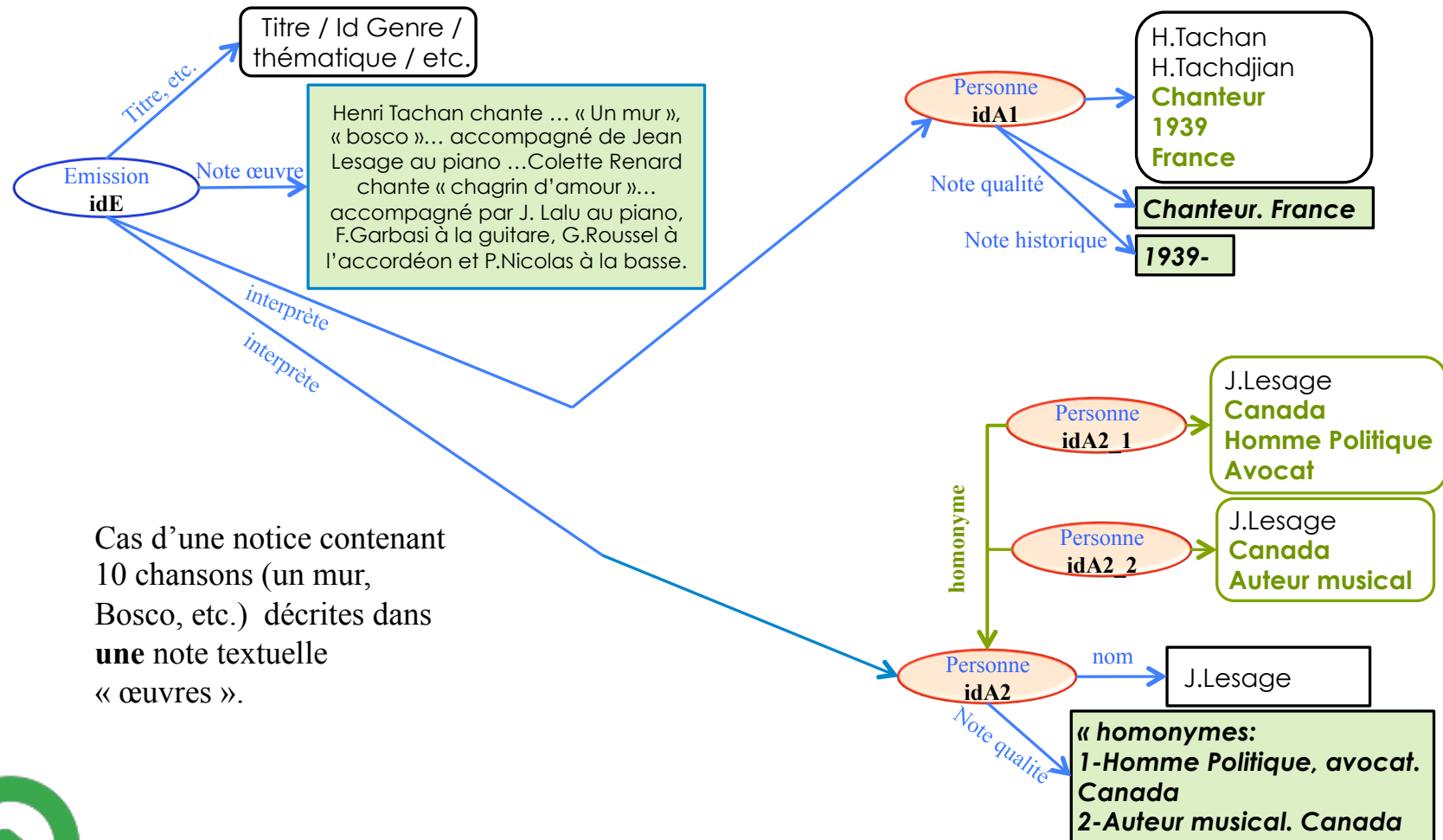
Traitement sur les Autorités

Extraction et Enrichissement sur notes textuelles:

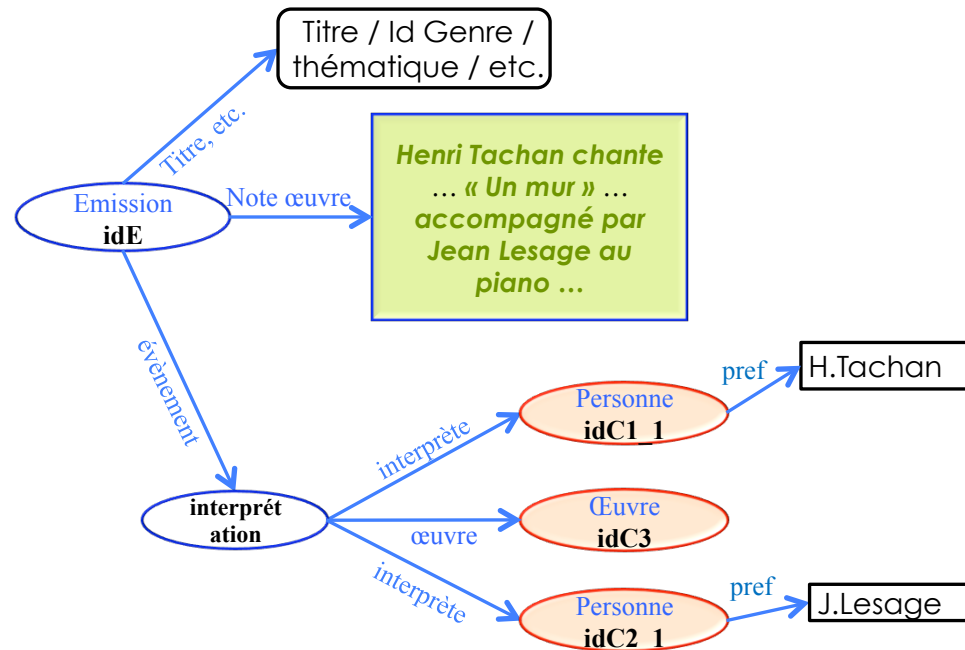
- Dates naissances / décès
- occupation(s) (avec métiers, organisation, lieux et période éventuellement)
- Nationalités
- ...



Traitement sur une notice de contenu

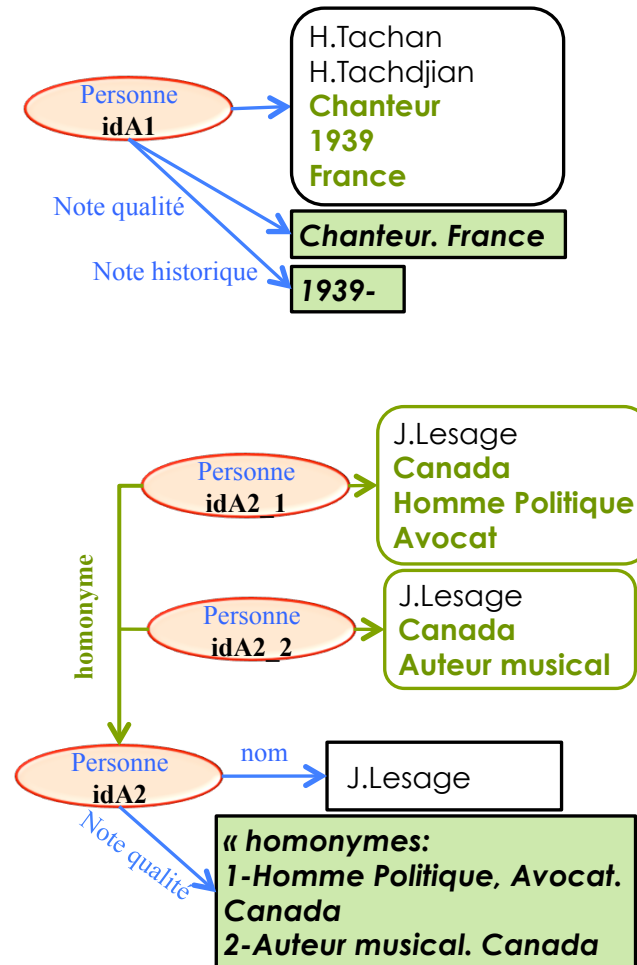


Référence et description contextuelle

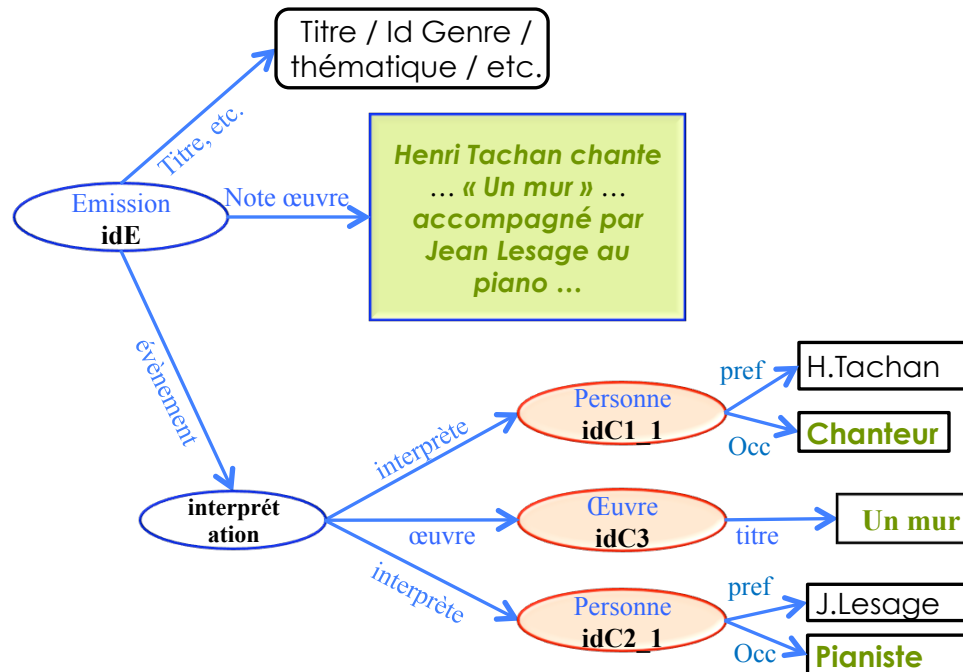


Extraction et enrichissement

Création du nœud « interprétation » (œuvre 1)
et des références contextuelles



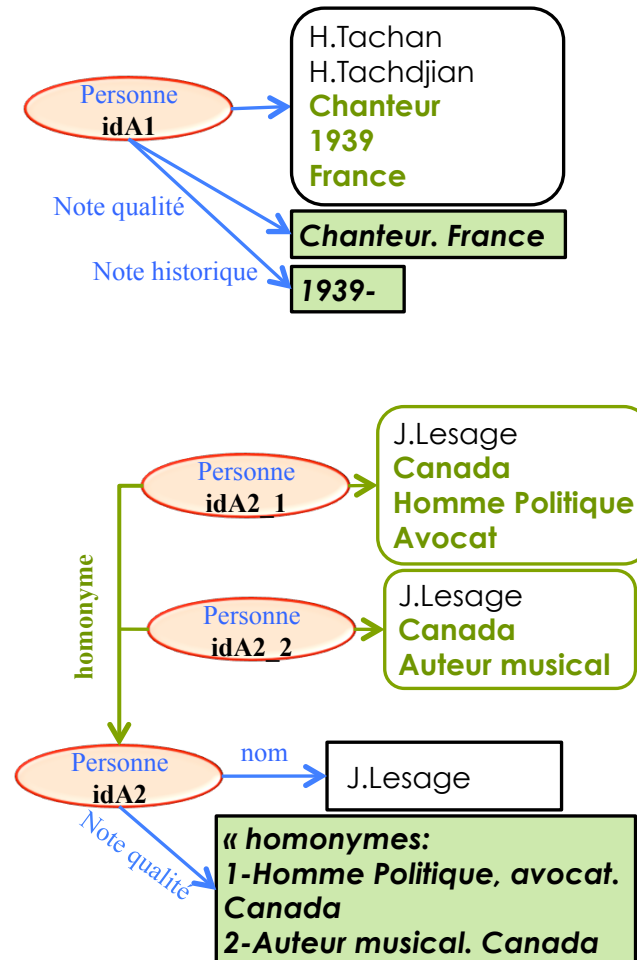
Référence et description contextuelle



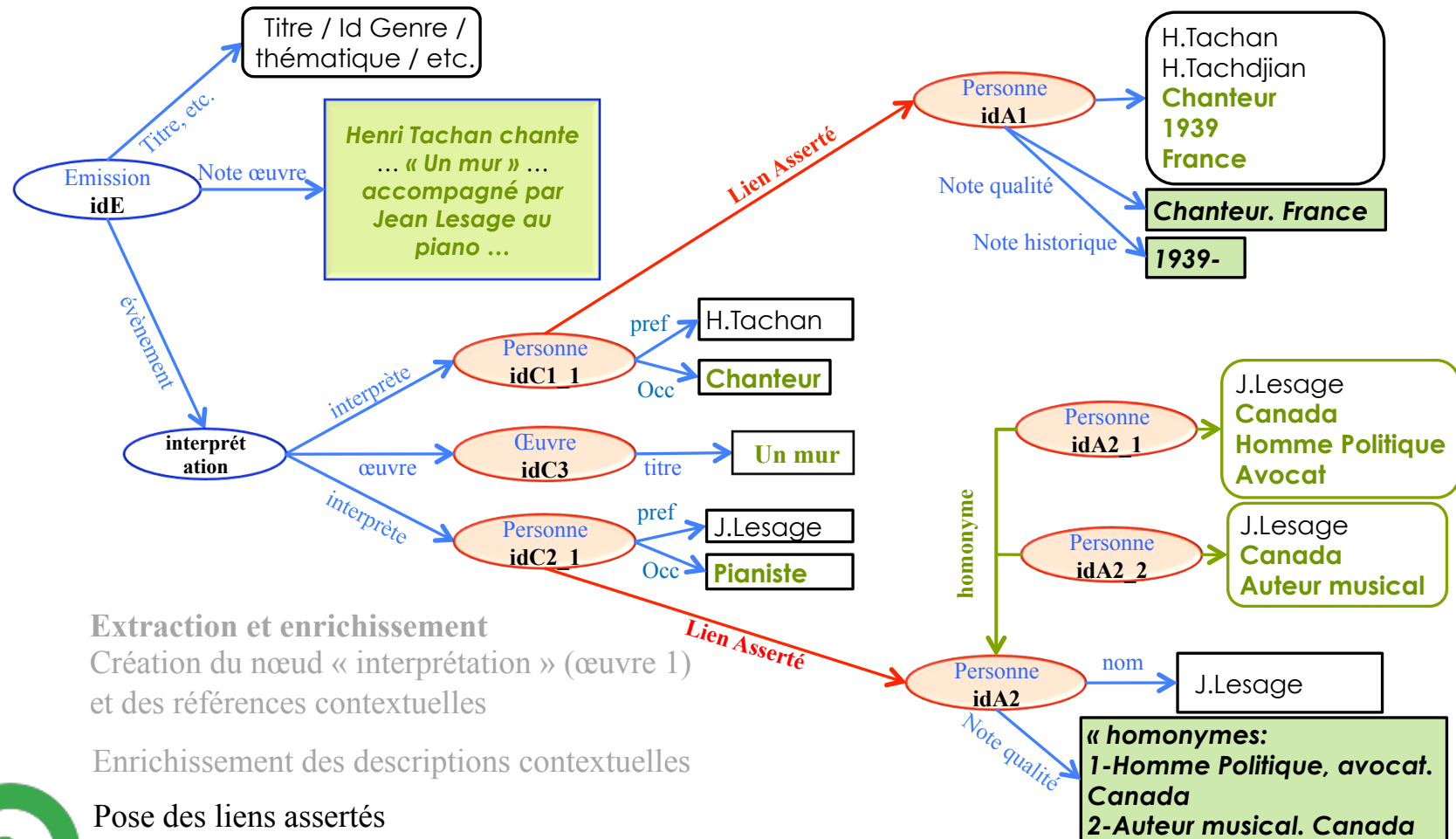
Extraction et enrichissement

Création du nœud « interprétation » (œuvre 1)
et des références contextuelles

Enrichissement des descriptions contextuelles



Référence et description contextuelle



Extraction et enrichissement
Création du nœud « interprétation » (œuvre 1)
et des références contextuelles

Enrichissement des descriptions contextuelles

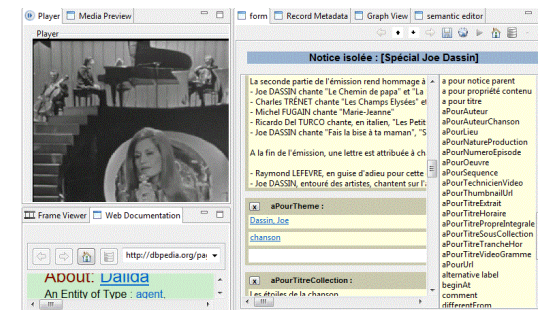
Pose des liens assertés

Réflexion sur les critères et règles associées

- Document de travail en cours avec les directions opérationnelles pour proposer des règles « métiers »
- Objectif: Soumettre des règles
 - pour guider l'extraction locale
 - *S'il existe un champ œuvre, dans une émission de thématique littérature et de genre magazine alors l'œuvre est très probablement un « livre ».*
 - Pour aider à la réconciliation contextuelle : dans un contexte local, entre contextes locaux, vers une « autorité Ina », ou vers une autorité du LOD
 - *Exemple: Si: même label de nom et prénom, même occupation dans organisation de même nom alors les personnes sont très probablement identiques*

L'utilisation prévue à l'Ina

- Un outil supervisant les corrections proposées (calculées off-line) dans un contexte de reprise d'antériorité (encours d'étude)
- Un outils d'indexation pour afficher et éditer des notices (documentaires et autorités) RDFS/OWL, pour évaluer l'aide apporté aux documentalistes



The screenshot displays a software interface with several components:

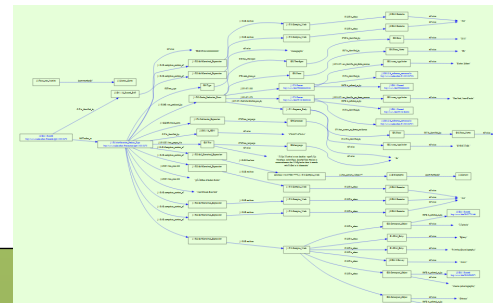
- Player:** A video player showing a black and white scene of a stage performance with musicians and a pianist.
- Media Preview:** A tab next to the Player.
- Form:** A tab next to Record Metadata, Graph View, and semantic editor.
- Record Metadata:** A section titled "Notice isolée : [Spécial Joe Dassin]" containing text about the second part of an emission and a list of artists.
- Graph View:** A section titled "aPourTheme :" with a list of items including "Dassin, Joe" and "chanson".
- semantic editor:** A section titled "aPourTitreCollection :" with a list of items including "Les étoiles de la chanson".
- Frame Viewer:** A tab next to Web Documentation.
- Web Documentation:** A section showing a URL "http://dbpedia.org/pa:" and a snippet of text: "ABOUT: Dalida An Entity of Type : agent."

La problématique ABES

Données. Vocabulaires

- Sudoc : 14 Millions NB / 2 Millions NA
- Format de travail : MARC (1960')
- Modélisation RDF pour Qualinca :
 - CRM CIDOC ISO 21127:2006
 - Extension FRBRoo
 - Extensions Qualinca
- Transformation (XSLT)

Collection: 4940-08-02-11-00012923011-11-19-30-30-11-2014
001 044-15-19-20-24-30-30-31-33-30-31-34-15-47-87-58-61-67-70-71-73-83-85-135-136-137-138-139-140-141-142-143-144-145-146-147-148-149-150-151-152-153-154-155-156-157-158-159-160-161-162-163-164-165-166-167-168-169-170-171-172-173-174-175-176-177-178-179-180-181-182-183-184-185-186-187-188-189-190-191-192-193-194-195-196-197-198-199-200-201-202-203-204-205-206-207-208-209-210-211-212-213-214-215-216-217-218-219-220-221-222-223-224-225-226-227-228-229-230-231-232-233-234-235-236-237-238-239-240-241-242-243-244-245-246-247-248-249-250-251-252-253-254-255-256-257-258-259-260-261-262-263-264-265-266-267-268-269-270-271-272-273-274-275-276-277-278-279-280-281-282-283-284-285-286-287-288-289-290-291-292-293-294-295-296-297-298-299-300-301-302-303-304-305-306-307-308-309-310-311-312-313-314-315-316-317-318-319-320-321-322-323-324-325-326-327-328-329-330-331-332-333-334-335-336-337-338-339-340-341-342-343-344-345-346-347-348-349-350-351-352-353-354-355-356-357-358-359-360-361-362-363-364-365-366-367-368-369-370-371-372-373-374-375-376-377-378-379-380-381-382-383-384-385-386-387-388-389-390-391-392-393-394-395-396-397-398-399-400-401-402-403-404-405-406-407-408-409-410-411-412-413-414-415-416-417-418-419-420-421-422-423-424-425-426-427-428-429-430-431-432-433-434-435-436-437-438-439-440-441-442-443-444-445-446-447-448-449-450-451-452-453-454-455-456-457-458-459-460-461-462-463-464-465-466-467-468-469-470-471-472-473-474-475-476-477-478-479-480-481-482-483-484-485-486-487-488-489-490-491-492-493-494-495-496-497-498-499-500-501-502-503-504-505-506-507-508-509-510-511-512-513-514-515-516-517-518-519-520-521-522-523-524-525-526-527-528-529-530-531-532-533-534-535-536-537-538-539-540-541-542-543-544-545-546-547-548-549-550-551-552-553-554-555-556-557-558-559-560-561-562-563-564-565-566-567-568-569-570-571-572-573-574-575-576-577-578-579-580-581-582-583-584-585-586-587-588-589-590-591-592-593-594-595-596-597-598-599-600-601-602-603-604-605-606-607-608-609-610-611-612-613-614-615-616-617-618-619-620-621-622-623-624-625-626-627-628-629-630-631-632-633-634-635-636-637-638-639-640-641-642-643-644-645-646-647-648-649-650-651-652-653-654-655-656-657-658-659-660-661-662-663-664-665-666-667-668-669-670-671-672-673-674-675-676-677-678-679-680-681-682-683-684-685-686-687-688-689-690-691-692-693-694-695-696-697-698-699-700-701-702-703-704-705-706-707-708-709-710-711-712-713-714-715-716-717-718-719-720-721-722-723-724-725-726-727-728-729-730-731-732-733-734-735-736-737-738-739-740-741-742-743-744-745-746-747-748-749-750-751-752-753-754-755-756-757-758-759-760-761-762-763-764-765-766-767-768-769-770-771-772-773-774-775-776-777-778-779-780-781-782-783-784-785-786-787-788-789-790-791-792-793-794-795-796-797-798-799-800-801-802-803-804-805-806-807-808-809-810-811-812-813-814-815-816-817-818-819-820-821-822-823-824-825-826-827-828-829-830-831-832-833-834-835-836-837-838-839-840-841-842-843-844-845-846-847-848-849-850-851-852-853-854-855-856-857-858-859-860-861-862-863-864-865-866-867-868-869-870-871-872-873-874-875-876-877-878-879-880-881-882-883-884-885-886-887-888-889-890-891-892-893-894-895-896-897-898-899-900-901-902-903-904-905-906-907-908-909-910-911-912-913-914-915-916-917-918-919-920-921-922-923-924-925-926-927-928-929-930-931-932-933-934-935-936-937-938-939-940-941-942-943-944-945-946-947-948-949-950-951-952-953-954-955-956-957-958-959-960-961-962-963-964-965-966-967-968-969-970-971-972-973-974-975-976-977-978-979-980-981-982-983-984-985-986-987-988-989-990-991-992-993-994-995-996-997-998-999-1000-1001-1002-1003-1004-1005-1006-1007-1008-1009-1010-1011-1012-1013-1014-1015-1016-1017-1018-1019-1020-1021-1022-1023-1024-1025-1026-1027-1028-1029-1030-1031-1032-1033-1034-1035-1036-1037-1038-1039-1040-1041-1042-1043-1044-1045-1046-1047-1048-1049-1050-1051-1052-1053-1054-1055-1056-1057-1058-1059-1060-1061-1062-1063-1064-1065-1066-1067-1068-1069-1070-1071-1072-1073-1074-1075-1076-1077-1078-1079-1080-1081-1082-1083-1084-1085-1086-1087-1088-1089-1090-1091-1092-1093-1094-1095-1096-1097-1098-1099-1100-1101-

[illegible]

Critères : fonctions de comparaison

- Nom du contributeur
- Type du contributeur
- Dates de vie du contributeur
- Période de publication
- Rôle du contributeur (en relation au document)
- Nature du document
- Langue
- Nom du co-contributeur
- Rôle du co-contributeur
- Éditeur
- Pays de publication
- Collection éditoriale
- Analyse du contenu : Dewey (comparaison statistique / sémantique)
- Analyse du contenu : RAMEAU (comparaison statistique / sémantique)

+ Fonctions de comparaison qui mêlent plusieurs attributs (ex : analyse de contenu à comparer ssi certains rôles)