Collaborative Repository for Research Corpora of Linked Digital Objects: RÉCOLTE

Elias Rizkallah[1], Ahmed Halioui[2]

[1] Département de sociologie, Centre d'ATO, Université du Québec à Montréal + Montréal – Canada

[2] Département d'informatique, Université du Québec à Montréal + Montréal – Canada

Abstract

This text exposes the rational, theoretic and main functionalities of a Web application to a collaborative repository for research corpora: RECOLTE. Steeped in the norms of collaboration between researchers and corpora elements (i.e. digital objects) enriched with metadata and inter-relations for reusability purposes, the proposed application is based on RDF linked digital objects, Dublin Core descriptions (extended) and XML-TEI exchange format (to import/export data). The architecture of the system, its access control management module for open and secured collaborations, massive/single import and export of data, search and navigate in the graph of objects, are all described and justified here within the context of logometric research.

Key words : linked digital object, collaborative corpus constitution, reusability, interoperability, RDFS, XML-TEI, logometric/textometric studies, online platform

1. Context[1]

The original idea of this Web-application was a paper entitled "Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche"[2] (Daoust, Duchastel, Marcoux, Rizkallah, 2008) where its goal was to propose a model for corpus deconstruction-reconstruction in the field of computer-assisted textual analysis. This paper was preceded by another one (Daoust, Marcoux, 2006) aimed to establish an XML-TEI exchange format between different logometric softwares. The RECOLTE-ATO[3] (hereafter shortened to RECOLTE) web-based platform represents an outcome of these preliminary stages in what could be called collaborative logometric (Mayaffre, 2005) or textometric studies. In the present section, the methodological problems that the application tries to solve are presented. Then, the theoretical inputs of Linked Digital Objects (LDO) will be exposed and finally the main functionalities of the application will be described through an exemplar corpus of LDOs.

---

[1] We wish to thank Mr. François Daoust for his multiple constructive comments on the manuscript of this text. In addition, the development of the online application has been funded by the Canada Foundation for Innovation (Project: 24266) and supervised by CRIM (Centre de Recherche Informatique de Montréal).

[2] Tentative translation: "for a repository model adapted to research corpus constitution"

[3] RECOLTE-ATO pour dépôt collaboratif de corpus de recherche d'objet numériques liés du centre d'ATO (Analyse de texte assistée par ordinateur). The url of the application is: http://entrepot.ato.uqam.ca/mcdapp. This application is currently in a Beta version and thus has few users so far.

In general, each logometric study addresses its research question by first thoroughly constituting a research corpus according to multiple criteria (*e.g.* relevance, representativeness, homogeneity) that globally tries to manage the unity of production-reception conditions, discourse type/genre and intertextuality. In any case, a corpus consists of several document units - hereafter called digital objects (DO) (Kahn & Wilensky, 2006) - that the researcher, before beginning any analysis, strives by thoroughly describing their properties, now commonly named metadata. Yet, during the same research project, each DO, apart from its relationship with the research subject, maintain different types of relationships with: itself (*e.g.* different versions), other DOs of the corpus (*e.g.* dependency, hierarchical) and other DOs outside of the corpus (*e.g.* DOs that at one stage belonged to the corpus). The first problem addressed by the application is thus managing metadata (descriptive and administrative) of DOs and mostly managing and representing all their semantic relationships, particularly since those procedures take place usually outside a logometric analysis software. The second problem concerns the collaboration between researchers whether from the same research team or of different ones. In fact, at a local level, a research project brings together different persons (professors, students, research associates, *etc.*) working in a team on DOs before, during and after the analysis stage. For instance, after the end of a logometric project, an actor in a research team could initiate a new project, thus constituting a new corpus the ingredients of which are already stored DOs. For that purpose he needs to access their descriptions, interrelations, analytical tools (codebook, codings, *etc.*) and analyses outputs (term frequency, characteristic terms, co-occurrence index, *etc.*). This point meets a re-usability issue, commonly treated nowadays in the field of "research data" (Borgman, 2010). On a more general level, the research community goes beyond the projects and collections of specific and known teams. The collaboration issue reaches here another level, the one of mutual enrichments between several research teams, and so beyond traditional ways (personal messages, meetings) through an online platform where actions (*e.g.* external annotations, addition of new documents) performed right on the data themselves, *i.e.* DOs. In short, the platform tries to meet needs of intra- and inter- digital objects relationships on one hand, and relations between several actors about research corpora, on the other hand, all under the mark of reusability by rich descriptions.

Before going more explicitly, here is an overview of the top 10 functionalities of the platform.

- Represent the semantic relations (incoming and outgoing) between DOs: hierarchical relationships (generic and partitive), versioning, annotation, dependency and derivation.
- Organize resources as a collection of objects (an object can belong to multiple collections and an object can contain multiple files).
- Manage multiple roles of actors (users, groups) having different rights in the system: Team Manager, Member of several teams, External user (human or Machine), etc.
- Manage accessibility of DOs: no access, restricted, complete, reading, editing.
- Submit queries (simple and advanced) in metadata and full text.
- Browse by facets within query results, and browse by network relationships through graphs between DOs.

- Add relationships between two DOs through the Graphical User Interface for crowd enriching the semantic networks between corpora.
- Batch description of multiple DOs during ingestion or a posteriori.
- Import one or multiple DOs with their descriptions (DOs and their relationships to other DOs).
- Export one or multiple DOs in an interoperable format (*i.e.* XML-TEI) with or without all of their relationships to other DOs.

Not all these functionalities will be described in section 3, only a group of the most relevant ones for logometric studies, but for now, the theoretical core of the application, *i.e.* the rich description (metadata and relationships) of corpora through linked digital objects must be exposed and justified.

2. Linked Digital Objects

Semantic Web technologies (Berners-Lee, Hendler, Lassila, 2001) offer essential functions for using computational methodologies in humanities. The Semantic web is the subject that provides the strategy, process and technology to share information in the Web. "*The Semantic Web was initiated [...] with an ambitious plan regarding the sharing of metadata and knowledge in the Web, enhanced with reasoning services for advanced new applications"* (Tim Berners-Lee). This new generation of services helps to assist humans in their problem-solving tasks, for instance, in searching information dispersed in texts, in a more categorized and structured Web. For example, looking for which Canadian minister was in charge during the war of Vietnam in the current Web 2.0 is a very laborious task. One should look for multiple web pages to build an answer. Categorizing key concepts and relations within and between texts should help to structure more relevant responses to complex queries. Formalizing queries to comprehend the goal of search and enhance the quality of results is an important step to represent the significance of terms (semantics) used in a search. In order to represent semantics grounded in queries elements, semantic Web uses the concept of triplets (or triples) of resources (Beckett & McBride, 2004).

Linked Digital Objects (LDO) (Tillett, 2004) is a semantic Web formalism to represent the semantics between resources (DOs) using the Resource Description Framework (RDF) language[4]. LDO descriptions use triplets of resources Uniform Resource Identifiers (URI) to represent statements. For example, in order to represent the annotation relation between two documents: *A isAnnotationOf B*, a triplet is generated as follows: *<N1:A> <N2:isAnnotationOf> <N1:B>*. A triplet represents a <subject> <predicate> <object> relation. Each triplet element represents a URI which is composed from a namespace and a name[5].

Due to their relevance in logometric studies, three types of relationships between digital objects will be presented here: (1) semantic relationships, (2) hierarchical relationships and (3) administrative relationships. The first kind of relationships describes semantically related

[4] https://www.w3.org/RDF/

[5] A namespace describes the controlled vocabulary (schema) of resources. The URI name represents the local name of one resource.

entities, *i.e.* DOs. We find in this category the annotation relationship between objects (*e.g. DO X annotates DO Y*) and other semantic dependencies, for example in a study corpus of financial reports, there is each year a dependency relation (*isDependentOf*) between each trimestral report and the final one. Hierarchical relations describe partitive (mereological or part-whole relations) and generic relations (as in set theory) between objects in order to structure them in a hierarchy of classes and subclasses. For instance, in a study corpus of books each chapter is in a merelogical (*isPartOf*) relation with the book from which it originates, whereas in a sub-corpus of press articles from a daily journal, each article is in a set relationship (*isMemberOf*) with all the other articles of that same daily journal. Finally, administrative relationships serve to manage physical resources in the repository. For example, the versioning relation (*DO X is a version of DO Y*) may be very useful to differentiate between corpus items (*i.e.* files) at time *t* vs time *t+1* while the derivation relation seem very useful between an image and its thumbnail.

As an illustration of the concepts of digital objects and triplets in logometric studies, we will take the "Nouvelle Édition du Comité d'Instruction Publique (CIP) des assemblées révolutionnaires en France" as a test corpus. Committee of Public Instruction (CIP) texts describe a new edition of minute recording sessions of revolutionary assemblies held from 1791 to 1793 in France. This new edition produced by Ayoub and Grenon (1997) presents 6354 pages composed from three sedimentary layers: (1) event documents (texts of minute recordings), (2) appendices and (3) annotation documents. These latter are originally commented by James Guillaume in 1889 and afterwards by Josiane Ayoub and Michel Grenon in 1997. The annotations of Ayoub and Grenon (*ag*) are actually annotations over the annotations of Guillaume (*ng*) which complicates the design of such corpus. Since CIP text corpus presents different kinds of semantic relations over different levels of abstraction, this corpus is used to test the representational and computational aspects of LDOs.

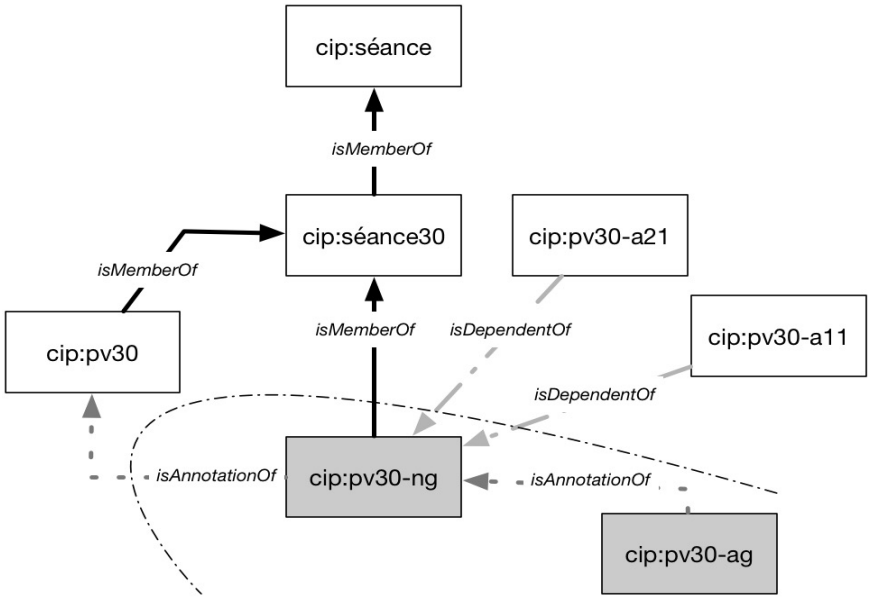

Figure 1. *Minute recordings of the session 30 (in Assemblée legislative). A session (séance) represents a set of hearing documents (pv30), appendices (a21, a11), annotations and comments (ng, ag) over these documents.*

Fig. 1 shows an example of a set of minute recording sessions triplets. Each digital object represents a content resource or an annotation one. Relations between these objects represent functional and semantic dependencies. For instance, the digital object of Guillaume (pv30-ng) annotates the minute recordings of the 30th session (pv30).

We use an ontology-based framework to organize and manage the graph of triplets. In order to validate and maintain described triplets, an RDF Schema is used. This latter offers a meta-conception about the structure of the ontology and hence, can handle more complex representations. For instance, a collection of objects, representing a logometric study subject (e.g. a research corpus), is a class of objects that could be part of many collections. A digital object could be used and shared in different collections with different users. Moreover, we can define a stack of levels of abstraction defining more complex annotations. Object relations could be transitive and, hence, we can find sequences of relations representing, for instance, annotations over already annotated digital objects. For example, the edition of Ayoube-Grenon text annotations is an edition over the original edition of Guillaume. Through the description logics of RDF and RDF Schema, different complex objects could be described. The syntax of these languages helps to represent a semantic network of digital objects. We adopt here the LDO formalism in order to represent text corpora.

3. RECOLTE-ATO Web platform

This section will first describe the system architecture of the platform, then explicit the multiple ways of populating it, later will be exposed the management of different access rights for collaboration and diffusion, and finally, navigation and search functionalities.

3.1. Project development and system architecture

The software development has been done according to SCRUM (Agile) methodology with great emphasis on documentation. The traces of the entire development process (decisions, implantations, customisations, versions and updates) are kept in a dedicated wiki (http://web.ato.uqam.ca/dokuwiki/doku.php). The final system architecture includes the following open source technologies:

- Fedora Commons (v 3.7): a warehouse of digital data systems supported by DuraSpace (http://www.duraspace.org);
- Hydra-Sufia (v 3.5): a collection of components facilitating "workflows" of digital data management in Fedora Commons (http://projecthydra.org)
- Apache Solr (v 4.3): a "sophisticated" full text search engine that uses a set of text analysis tools (https://lucene.apache.org/solr/)
- Blacklight (v 4.5): a "new generation" data search Web interface supported by its own community (http://projectblacklight.org)
- Solrizer (v 3.1.1): a component that plays an articulating role between the layer Fedora & Solr and the one of Hydra-Sufia by providing indexes and triplets to the latter.

The whole dynamic between the different components can be graphically represented:

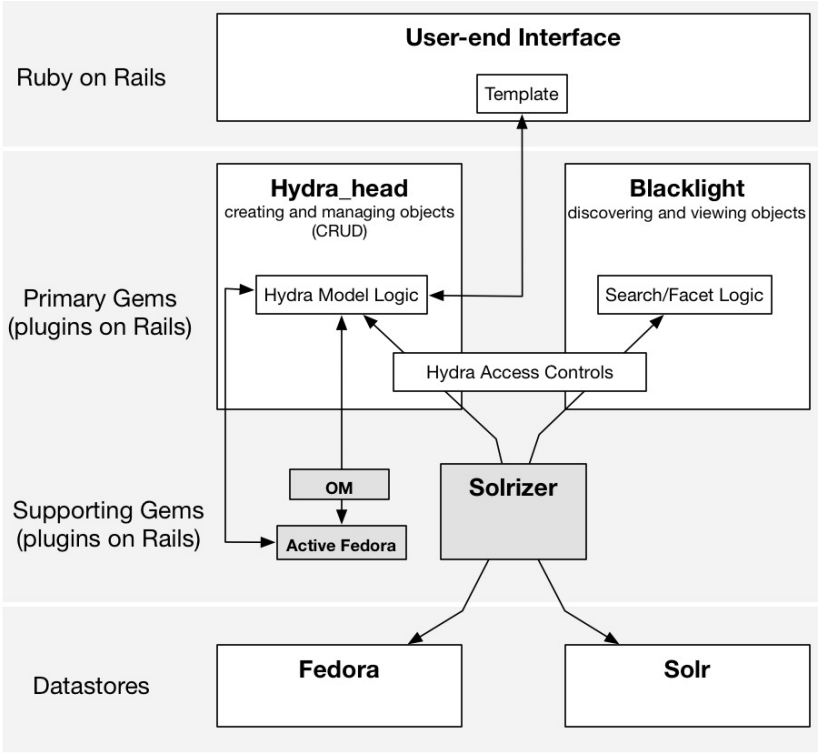

***Figure 2**. System architecture. An updated version from the Hydra Duraspace Website: https://lib.stanford.edu/libdevconx3/hydra-public-apis*

Datastore level describes the triplestore storage of DOs (Fedora) and their indexes (Solr). The second level presents a set of Ruby Gems[6] managing the DO logic model (Hydra-head) and the search module (Blacklight and Solrizer). The Hydra Access module handles access controls over DOs and manage the application level above Fedora. The end-user graphical interface (upper level) brings its templates from Hydra and its navigation tools (e.g. facets) from Blacklight. Lastly, even if the whole user interface is in French language, the system architecture is flexible enough to easily add another language.

3.2. Repository populating

We describe in this section the way DOs are stored in datastores and thereby the rich description issue for enhanced reusability. Fedora Commons uses a "compound digital object" design which aggregates several content items into the same DO. Content items or datastreams, as Fedora call them, represent either data or metadata. Our DO logic model defines particular contents to store the research text corpora: (1) RELS-EXT (Relationships External), (2) DC (Dublin Core, DCTERMS and MARCRel), (3) RIGHTS and (4) CONTENT datastreams. RELS-EXT datastream is used to describe relations between objects in RDF. DC datastream is used to contain metadata describing the object. In order to enhance the precise description of each DO, and hence its reusability, to the traditional 15 elements of the Dublin core, we added two types of refinements: a) the MARCRel[7] roles of each

[6] Ruby gems are external Ruby programs and libraries. They contain package information along with files to install.

[7] http://dublincore.org/usage/documents/relators/

contributor (*e.g.* editor=edt, annotator=ann, programmer=pgm); b) two DCterms[8], one for the temporal coverage of the object (*e.g.* 18th century) and the other for its spatial coverage (*e.g.* France). Administrative metadata are automatically added while uploading an object (*e.g.* MIME type, creation date, *etc.*). The RIGHTS datastream stores the control access data to Fedora and Solr datastores (*cf.* section 3.3). Default policies restrict by default any access. This can be changed while importing DOs by the repository. And finally, the CONTENT datastream stores data (*i.e.* text) that constitutes the DO. The following figure shows in an RDF file these four levels of a DO.


```
<?xml version="1.0"?>
<rdf:RDF xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:marcrel="http://www.loc.gov/marc.relators/"    >

  <rdf:Description rdf:about="2738457916-pv031-ng.xml">

    <rdf:master_collection>EVR, législative</rdf:master_collection>

    <dc:title xml:lang="fr">Notes de J. Guillaume sur le procès verbal de la séance 31 du
    <dc:creator>Boulad-Ayoub, Josiane</dc:creator>
    <dc:language>fr</dc:language>
    <dc:publisher xml:lang="fr">UQAM, Projet d'encyclopédie virtuelle des révolutions</dc:
    <dc:rights>http://creativecommons.org/licenses/by-nc-sa/2.5/ca/</dc:rights>
    <dc:source xml:lang="fr">France. Assemblée nationale législative (1791-1792). Comité d
    <dc:type>Text</dc:type>

    <dcterms:spatial xsi:type="dcterms:ISO3166">FR</dcterms:spatial>
    <dcterms:temporal xsi:type="dcterms:Period">start=1791-10-30; end=1792-09-22;</dcterms
    <dcterms:created>2012-09-28</dcterms:created>

    <marcrel:ann>Guillaume, James</marcrel:ann>
    <marcrel:edt>Boulad-Ayoub, Josiane</marcrel:edt>
    <marcrel:prg>Daoust, François</marcrel:prg>
    <marcrel:spn>CRSH du Canada</marcrel:spn>
    <marcrel:spn>Chaine UNESCO</marcrel:spn>
    <marcrel:spn>Université du Québec</marcrel:spn>
    <marcrel:spn>Chaire MCD, UQAM</marcrel:spn>

    <rdf:isMemberOf rdf:resource="2738457916-seance031.xml"/>
    <rdf:isAnnotationOf rdf:resource="2738457916-pv031"/>

    <rdf:text>31e séance 210. Voir le texte de ces articles au procès-verbal de la séance
    du Comité du 12 décembre 1791. Il avait été entendu précédemment que le travail de
    rédaction du projet de décret sur les congrégations séculières serait partagé entre
    le Comité des domaines et celui de l'instruction publique ... </rdf:text>

  </rdf:Description>
</rdf:RDF>
```

Figure 3. *Digital object RDF description file.*

In order to import DOs into our repository, we use the exchange format RDF XML to describe the object and XML-TEI (Daoust & Marcoux, 2006) to describe the content of the analyzed text. XML-TEI exchange format was conceived as a pivot format in order to enhance reusability by allowing corpora to be easily imported and exported between 4 common textometric softwares : Alceste, DTM-Vic, Lexico3 and SATO (for an application of its pivot power, cf. Daoust et al. 2006). The proposed XML RDF metadata file presents two sections (see *Fig.* 3). The first section is about DC, DC-TERMS and MARC-REL metadata annotating the resource with the Dublin Core vocabulary to describe the properties of analyzed texts. The second section of the XML file describes the RDF relationships between DOs [9]. Datastreams are generated automatically thereafter from the imported XML RDF and TEI files.

[8] http://dublincore.org/documents/dcmi-terms

[9] Please note that relations metadata are optional to upload as same as for some other metadata like spatial coverage or sponsors. More detailed information about our XML RDF format is available on our Wiki Website: http://web.ato.uqam.ca/dokuwiki/doku.php

3.2.1. Single import and export

The single import/export service allows users to upload and export one DO and its content(s). This can be used for exchanging documents with another repository, for archival or backup purposes. In response to a command to import/export selected document, the server invokes the Hydra module (DO model) which parse the object's metadata (in RDF) and its content (in XML-TEI). A specific RDF Schema validates the syntax and graph consistency of the imported DO. This procedure verifies if the DO is correctly formed and there is no semantic inconsistency within the graph of interrelations. For instance, the schema verifies if a DO is not part of itself. Next, the server uploads DO constituents in the several datastreams. Each stored datastream is treated as an opaque bit stream. It's up to users to export the desired content.

3.2.2. Batch import and export

This service allows users to exchange several objects that have been edited and enriched during the logometric analysis stage. Importing in batch uses the same exchange formats to upload a single DO. The massive import allows users to upload their research corpus DOs and edit a meta-document which describes the import process. A meta-document is Web interface allowing to describe the corpus, its constituents and control access to them. Users could also export a large amount of DOs through the search interface. Selected objects are sent to a specific service in order to download their datastreams and verify if users are allowed to access them.

3.3. Access rights, collaboration and diffusion

RECOLE-ATO favours collaborations between researchers by installing a fully fledged rights management functionalities which allows both to collaborate with different members of research teams while maintaining control over data accessibility and possible actions on 4 levels: DOs, other individual actors, relationships and groups of actors (*i.e.* research team).

The following table synthesizes the different type of users and their permissions on these 4 levels.

			Role / User type					
			Non subscribed	Subscribed not member of any group	User member of a group	Group adminstrator	Research Center or collection director	Super Admin
Location of permissions	Digital Objects	Upload			X	X	X	X
		Allow upload				X	X	X
		Submit to publication				X	X	X
		Publish					X	X
		Modifiy			X	X	X	X
		Download a document from group			X	X	X	X
		Add to Favorites		X	X	X	X	X
		Download public DO	X	X	X	X	X	X
	Relations	Add relations beween 2 public DOs		X	X	X	X	X
		Delete relations beween 2 public DOs					X	X
	Other Users	Add			X	X	X	X
		Delete				X	X	X
		See			X	X	X	X
	Group	Create					X	X
		Delete						X
		Modifiy				X	X	X
		Give permission to create						X

Figure 4. *Control access policies.*

In this figure, one could notice that the span of rights according to role/user type increases incrementally from left (non subscribed, i.e. anonymous user) to right (Super Admin). Thus, a subscribed user, not belonging to any group, can only download/export DOs of a public corpus, add DOs and add relations. Whereas a group administrator is allowed to do the same as the previous one, in addition to having all rights on his DOs, other users of his group (add/ delete/display/allow upload) and the configuration of his group.

On the other side of the coin, the actor uploading one or multiple DOs can determine the accessibility of the object: no access, restricted (metadata only), complete, editing (relations). Hence, if "restricted" is selected, the actor allows the community to know about what his team is working on without giving full access to the files associated to the DOs of the project. Consequently, the management of access facilitates team work, diffusion of projects, and mostly the reusability of DOs between research teams especially that import-export format (XML-TEI) allows analysis in different logometric softwares.

3.4. Navigate and search

We use the SPARQL 1.0 query language[10] to navigate through the labeled, directed RDF graph. SPARQL queries are sent from a client to a service, i.e. search using the HTTP protocol. This offers to users the possibility to query the corpora from different perspectives and facets. SPARQL is specifically used to find object occurrences using RDF triplets and regex patterns. With using SPARQL and RDF expressiveness, retrieving objects in a complex directed acyclic graph tend to be a straightforward task. SPARQL provides a powerful language to query operations such as OPTIONAL (join) and AGGREGATE. For example finding the objects annotating the session N°30 created after April 2015 is expressed using the RDF expressions to formulate the annotation relation and a regex pattern on the administrative metadata: date of creation. However, using a regex is a very expensive operation. Besides of other SPARQL 1.0 limitations such as the negation, constraints on graph paths (reflexivity, transitivity, etc.), we choose to use the full text search engine Solr[11].

Solr enables advanced full-text search capabilities including indexing, matching phrases, wildcards and many linguistic analysis tools such as tokenization, stemming, etc. We use Solr to index metadata values for fast information retrieval. For example, in order to retrieve the documents talking about triumphal pumps in minute recordings, it is sufficient to find the index of the stems "triumph" and "pump'. Solr enables also faceted search which play a key role in our application. Faceted search addresses weakness of conventional search approaches and proved a more intuitive information retrieval support to users. Several metadata are faceted. Through a multi-select facets, one can use conjunctive metadata values in order to refine a search. A sequence of refinements laid out in search query could be used to retrieve the wished objects(s) in the graph of DOs.

A human-friendly interface to navigate through the graph of objects is offered. A partial graphical form of the graph is presented while accessing a DO. The user could navigate

[10] https://www.w3.org/TR/rdf-sparql-query/

[11] http://lucene.apache.org/solr/

through the semantic graph by choosing the level of the desired ontological abstraction and the types of relationships outgoing the concerned object. This gives a visual perspective of the DOs (see *Fig.* 5) and allows another kind of navigation through the semantic graph using SPARQL. Querying the graph of documents is used in order to retrieve digital objects. Yet, analysing the morpho-syntactic features of object contents is dealt with Solr engine. Hence, with coupling the semantic power of SPARQL and the analytical power of the full-text search of Solr, we implemented a robust search and navigate engine to retrieve the most relevant results for a query.

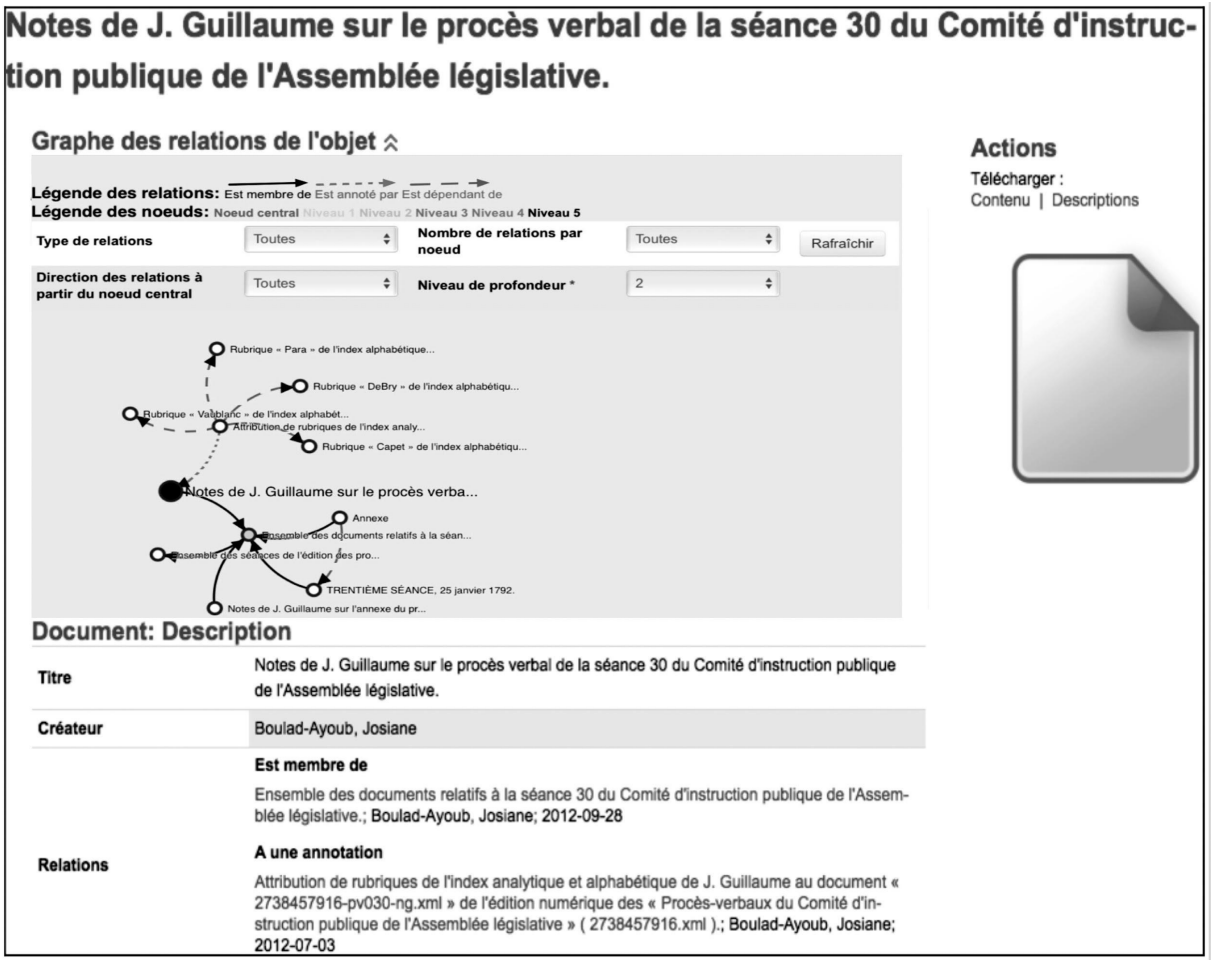

Figure 5. *Digital object view interface in RECOLTE:*
http://entrepot.ato.uqam.ca/mcdapp/files/v979vr35d

4. Conclusion

As exposed, RECOLTE-ATO application promises interesting potentials for the logometric studies community by including several features and different schemas for a rich description of digital objects and their inter-relations; while at the same time favouring secure storage of data, broadcasting logometric projects, team working, and mostly making profit of the reusability of corpus items between projects. Though, several limitations should be noted and extensions are to be expected. Regarding limitations, one should note that SPARQL is only used in the backend of the application and customized queries using SPARQL are only possible through FEDORA GUI. SPARQL queries could be very useful to query complex

semantic relations (*e.g.* retrieving all DOs that are on several hierarchical levels (members, sub-members, etc.) form a given vertex in a graph of DOs and thus making a full use of SPARQL expressiveness capabilities. Although storing objects in RDF triples allows many flexibilities in operations, batch uploading a high amount of objects (> 1500) could be time consuming in terms of computation; a graph database (*e.g.* Neo4j) communicating with Fedora could be a very promising path of solution. On the top of the list of our priorities in terms of extensions, there are: 1) to implement a major extension to RECOLTE in order to construct-deconstruct a corpus through a graphical interface in the application and not outside of it; 2) to enhance the number and nature of softwares (not only textometric one but also the custom annotation oriented) compatible with the XML-TEI exchange format in order to optimize the interoperability between researchers and common corpora.

5. References

Ayoub J. & Grenon M. (1997). *Édition nouvelle, présentée, mise à jour et augmentée des procès-verbaux du comité d'instruction publique*. Paris : L'Harmattan

Beckett, D., & McBride, B. (2004). RDF/XML syntax specification (revised). W3C recommendation, 10.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 28-37.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? (pp. 1–21). Presented at the China - North American Library Conference. Retrieved from http://www.ratswd.de/download/RatSWD_WP_2010/RatSWD_WP_161.pdf

Daoust, F., Dobrowolski, G., Dufresne, M., & Gélinas-Chebat, C. (2006). Analyse exploratoire d'entrevues de groupe: quand ALCESTE, DTM, LEXICO et SATO se donnent la main. In *Actes des 8èmes journées internationales d'Analyse statistique des Données Textuelles: JADT 2006* (p. 313-326).

Daoust, F., Duchastel, J., Marcoux, Y., & Rizkallah, É. (2008). Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche. *Actes 9e Journées internationales d'Analyse statistique des Données Textuelles*, *1*, 355–367.

Daoust, F., & Marcoux, Y. (2006). Logiciels d'analyse textuelle: vers un format XML-TEI pour l'échange de corpus annotés. In *8e Journées internationales d'Analyse statistique des Données Textuelles* (Vol. 1, p. 327-340).

Kahn, R., & Wilensky, R. (2006). A framework for distributed digital object services. *International Journal on Digital Libraries*, *6*(2), 115-123. http://doi.org/10.1007/s00799-005-0128-x

Mayaffre, D. (2005). Analyse du discours politique et Logométrie : point de vue pratique et théorique. *Langage et Societé*, (114), 91–121.

Tillett, B. B. (2004). Authority control: State of the art and new perspectives. Cataloging & Classification Quarterly, 38(3-4), 23-41.