FROM SEMANTIC WEB TO FACETED CLASSIFICATION
A CASE STUDY AND FIVE LINES OF FUTURE RESEARCH

Carlo Bianchini
Department of Musicology and Cultural Heritage, University of Pavia, Pavia, Italy

Keywords:
*Faceted classification, Semantic web, Colon Classification, **data.bnf.fr**, Italian Literature classification.*

Abstract

A case study about the automatic creation of faceted class numbers for special topics starting from data available on the semantic web is described. The aim is to investigate two questions: 1) Is the automatic, quick, cost-effective, large-scale production of a Schedule of Colon Classification class numbers for literary classics possible?, and 2) More generally: which are the requirements for an automatic production of class numbers of a faceted scheme of classification? What kind of research is necessary to develop such a process?
*The case study focuses on data extracted from **data.bnf.fr** and on their reuse to get correct and complete **Colon Classification** class numbers for a sample of Italian literature writers (born from 1800 to 1900). The process of creation of class numbers starting from data identification and extraction, through their cleanup, transformation and translation in classified notation by means of Open Refine is presented. Case study results and their relevance for starting questions are discussed and research questions for possible future developments are identified.*

Introduction

Ranganathan's desire to overcome existing classification schemes and to create a new one based on new semantic and syntactical features is the foundational idea of *Colon Classification* (CC) (Ranganathan, 1963). In the new classification scheme, simple and composed subjects[1] are to be expressed by means of semantic and syntactical tools represented in the facet formula. Ranganathan states this basic idea by the contraposition of schemes *of* classification (e.g. Enumerative Classifications) – and schemes *for* classification (e.g. Faceted Classifications) (Ranganathan, 1965, pp. 13-14).

Enumerative Classifications are founded on an articulated scheme that contains all the necessary numbers for the most parts of the subjects. Ranganathan explains that

> An Enumerative Scheme for classification consists essentially of a single schedule enumerating all subjects – of the past, the present and the anticipatable future.
> Such a schedule will have necessarily to be long. Further, it will soon be overpowered by the emergence of subjects beyond the anticipated ones – overpowered in the sense that it would be difficult to place each new subject in its filiatory position among the already existing subjects (Ranganathan, 1967).

In this sense, *Library of Congress Classification* (LCC) and Rider's *International Classification* are designed as Enumerative Classifications. *Dewey Decimal Classification* (DDC) is an Almost-Enumerative Scheme for Classification, as it has in addition a few schedules of common isolates (*Tables*). Stefano Tartaglia underlines that

> Enumeration implies either a deliberate limitation of the number of the classes and, in this case, it is functional just for a non-analytical classification, or the provision of every combination of concepts that might constitute the subject of a document; but the latter is impossible,

1 A simple subject is a subject constituted by one 'isolate' – that is one idea, that can be represented by one authorized term and one notation composed by one or more figures. For example, the subject of a book about *mathematics*, that is represented by the idea of Mathematics, expressed by the term *mathematics* and, in CC, by the figure "B" is a simple (or basic) subject. Instead, composed subjects are constituted by two or more isolates, and, depending on the kind of relationships among these isolates, can be compound or complex subject. For example, "History of mathematics in Europe during the XVII Century" and " Encyclopedia of Italian linguistics" are two composed subjects; on the contrary, "Statistics for librarians", and "Influences of Dante on Leopardi" are two complex (or phased) subjects (Ranganathan, 1959, 1967).

> so that the provision results in a selection – always partially aleatory – of the most probable subjects (Tartaglia, 1998, p. 12).[2]

CC is an analytic-synthetic classification, and it does not supply ready-made classification numbers, but *Standard schedules* that provide a list of 'isolates' (i.e., the base elements of the 'Meccano' that constitutes the classification) to be used to compose a class number representing the simple or composed *specific subject* of a resource.[3] So, if the former type of classifications is based on an aleatory anticipated prevision of any possible subject to be classified (and in this sense they are scheme *of* classification), faceted classifications provide just the basic tools to compose each time a number containing all the necessary isolates (so that they are a scheme *for* classification). By a different process of synthesis based on the numbers listed in the schedules, class numbers for any simple or composed subject can be built by combinations and permutations allowed by syntax rules (such as the Facet Formula placed at the beginning of each Main Class Schedule). Ranganathan's intuition was confirmed by a series of attempts and experiments, that led to the conclusion that:

> a Scheme for Classification, suited to books embodying multi-faceted subjects, should provide for two or more free facets and not merely one. This was the predisposing cause for designing a faceted Scheme for Classification (Ranganathan, 1965, pp. 13–14).

But there are exceptions to the rule! As well as in DDC the organization is not completely enumerative (Tartaglia, 1998, p. 7; 18-19), in CC the organization is not completely an analytic-synthetic one: for practical reasons – and for coherence and uniformity sake – a part of the three folded text of the CC, 6th edition (Ranganathan, 1963), is a scheme *of* classification. In fact, Part 3 contains Schedules of Classics and Sacred Books with special names, i.e. a list of ready-made class numbers for classic works in many main classes such as *L Medicine*, *Δ Spiritual Experience and mysticism*, *N Fine arts*, *O Literature*, *P Linguistics*, *Q Religion* and *R6 Indian Philosophy*.

These enumerative schedules of classics are an extremely useful tool, especially for Indian libraries. Nevertheless, it is very difficult to find some class numbers interesting for a Western library among the more than 126 pages of Part 3 of the

2 Translation from Italian by the author.

3 In *Elements of library classification*, the specific subject of a book is defined as "that division of knowledge whose extension and intension are equal to those of its thought-content" (Ranganathan, 1959, p. 4).

CC. If a Part 4 of CC devoted to Western classics – or at least literary classics – were available, the use of CC would be simpler, more efficient and more effective.

Nevertheless, the construction of this hypothetical Part 4 of CC would be a considerably time-consuming work, additionally to be done without any warrant that efforts would be rewarded by an actual and large use of new Schedules. Therefore, a question arises: is the automatic, quick, and cost-effective large-scale production of a Schedule of class numbers for Western literary classics possible? Moreover, the answer to this question implies considering a more general issue: which are the theorical requirements for an automatic large-scale production of class numbers of a faceted scheme of classification?

Two prerequisites seem to be essential to answer in the affirmative both questions: 1) creation of class numbers must be a regular, exception-free and predictable process, so that it could be transformed into an algorithm; 2) modular, shareble, reusable and preferably free data needed to create facets provided for in the classification scheme must be available. The former prerequisite is granted by two specific characteristics of CC, i.e., the availability of the *Classic device* and the semantic approach provided by facet formula in Main Class *O Literature*. Both the characteristics allow defining an algorithm for creating new class numbers. The latter prerequisite is theoretically granted by Linked Open Data (LOD) available on the Semantic Web.

Are these prerequisites enough? The case study on *data.bnf.fr* and CC aims to develop the basis to evaluate the opportunity of a more extensive research on how LOD might be used for subject indexing by faceted classification schemes.

Classic device and semantics of Main Class *O Literature*

In CC, the classification process for literary works can be transformed in an algorithm thanks to classic device and facet formula of the Main Class *O Literature*. In CC, a classic work is 'a book stimulating other books and literature on itself'; it is also defined as a 'work (other than a Sacred Work or a Work of Literature) expounding some specialised subject, usually having embodiments in several versions, adaptations, and translations, inspiring other works on itself, and getting copied out and/or brought out in print even long after its origin' (Ranganathan, 1963, p. 1.59, 1967, p. 486).[4] So, classic works exist in Literature, of course (e.g., Homer's *Iliad*), but also in any other discipline: Darwin's *The origin of the species*,

4 The definition is very clear and useful also in the IFLA LRM perspective, as it implies that a work should be considered a classic every time it stimulates either new expressions, or manifestations, or items, or other related works.

Aristotle's *Logic*, and Charles A. Cutter's *Rules* are all examples of classic works, as they stimulated other books and literature on themselves. Ranganathan notes:

> This is strictly a classifier's definition. It naturally uses that quality of Classics which challenges a classificatory language to deal with it properly. It wants that a Classic and its associated literature should be arranged like a continuous spectrum, with nothing outside itself intervening. […] The inherent qualities of a Classic stimulating such auxiliary literature are that: 1. It has elements of permanent value; 2. It is saturated with the personality of the author – which in itself was very powerful and highly organized; and 3. It is a seminal book cutting new ground, blazing new trail, stimulating new thought, and so on.

In CC two types of classic *devices* are provided; the former is devoted only to main classes *O Literature* and *Q Religion* and is based on a chronological principle, since chronological organization is considered more useful for these two disciplines. The latter applies to any other Main Class, and takes advantage of a subject approach, and subordinately of a bibliographical-literary approach. For example, the application of the classic device to the Aristotle's *Organon* results in an organization (*filiatory sequence*) of its expressions and manifestations, and subsequently of its comments (organized by their expressions and manifestations too).[5]

Main class *O Literature* is probably a *unicum* in CC, as its Schedules for classification consists in a list of less than 10 isolates (voices); it is also interesting because two out of four facets (Language and Time) of its formula can be obtained from common isolates Schedules.[6] Facet formula of Main Class *O Literature* is the following (CC, 2.94):

O [P], [P2] [P3], [P4]

where facet [P] represents the language, facet [P2] corresponds to literary form (see below), facet [P3] stands for author (represented by means of the birth year), and facet [P4] corresponds to the work facet (CC, 1.101-1.102). Every facet is intended to be used only when it applies.

Overall, facet formula means that, in Main Class *O Literature*, concepts and

5 For an example of application of the classic device to Aristotle's *Organon* and of the organization of expressions and manifestations of the work and its comments, see (Bianchini, 2012; Varghese, 2010).

6 Common Isolates in CC have their rough correspondent in Auxiliary Table of DDC.

punctuation must respect the following order:

O + Language + , + Literary Form + Author + , + Specific Work

The first facet [P] is determined by the Language Isolate schedule; Language facet expresses the language of the work (in a FRBR meaning).[7] For instance, as isolate numbers for English, German, and Russian languages are respectively 111, 113, and 142, English literature gets class number O111, German Literature gets O113, and Russian Literature gets O142. Facet formulas show that facet [P] is separated from facet [P2] by means of a comma. The second facet [P] is defined by the isolates of the Main Class *O Literature*; the array is fully presented in the *O Literature* schedule, and it is faceted by the literary form, as follows:

1 Poetry
2 Drama
3 Fiction, including short stories
4 Letters (literature written is the form of letters)
5 Oration
6 Other forms of prose
7 Campu[8]

Facet [P2] stands for the literary form of the work; its focus is "the form of literature which the book contains or about which the book treats" (CC, 1.98). For instance, a book containing Russian poetry takes the class number O142,1 (i.e. a comma and the isolate number 1 for Poetry are added to O142). In the same way, a book about English Drama gets class number O111,2.

Up to this point, CC and DDC facets are very similar: in both, the sequence of ideas and notation express first main class Literature, then language of the work, and lastly literary form of the work. Though, from the third facet onwards, CC becomes more specific, precise, and accurate, because it allows identification of both the author and, eventually, each single work of the author.

In the facet [P3], the author of the literary work or subject of the book is to be represented by the birth year. Years are determined by means of the Time isolate schedule, where each century is represented by a capital letter. For instance, the

7 In fact, in CC the other FRBR Group 1 entities – such as Expression, Manifestation, and Item – are represented by attributes expressed within the book number; see below.

8 Champu: a literary form of the Indian literary tradition (Kannada and Telogu languages) that uses a blend of prose and poetry.

letter J represents all the years from 1500 to 1599, N represents the range 1900-1999 and P represents the range 2000-2099. Accordingly, each single year of any century can be obtained by adding its two last figures to the Letter representing its century, so that 1564=J64, 1956=N56, and 2001=P01.

For example, by means of Schedule of Time, a book about Shakespeare's plays gets the class number O111,2J64 (as Shakespeare was born in 1564). The author indexing system created by Ranganathan is modern, because with this faceted approach a literary author is almost always given a unique notation.[9] It could also be used as an author identifier to some extent, and it allows representing an author both as a creator and as a subject in the same way within Ranganathan's classified catalogue.[10]

The fourth facet [P4] refers to the work or the works contained in the book, or about which the book is. The most relevant situation for the case study is that of books containing one work. In fact, in this case the isolate number for facet [P4] is a progressive special number assigned by the classifier,[11] following the chronological sequence of the author's works.[12]

To sum up, the reason why it is possible to produce CC class numbers by an algorithm is a specific characteristic of the main Class *O Literature*. Additionally, every data element to be used as a value in its facet formula – language, literary form, the author's birth year and the title of the work – is already available either in bibliographic and authority records or as Linked Open Data. For this reason, an algorithm for the automatic construction of class numbers for Main Class *O Literature* can check the values of the relevant data elements and transform them in the corresponding isolate numbers to synthesize the final class number of the book.

Data elements to be provided for the facet formula are the following:

- author's nationality, to identify the relevant national literature;
- literary form mostly associated with the author;
- author's birth year;
- number of works associated with the author;

9 It must be noted that the obtained class number represents more exactly *any* dramatic English author born in 1564, as O111,2 represents English Drama, and J64 stands just for 1564 (a year associated with William Shakespeare, but theoretically associable with any other dramatic author born in that year). Anyway, rules for disambiguation of authors are provided in case of ambiguity (CC, 1.100-101).

10 For further readings about Ranganathan's Classified Catalogue, see (Ranganathan, 1964).

11 The integer must be assigned by a special process to work out the exact amount and value of the figures (depending on the number of the author's works, if it is not exceeding 8 works, or it exceeds 8 but not 64 works, or it exceeds 64 works; CC, 1.101-102). For the sake of brevity, and provided that the process is completely automatable, a simply progressive integer was used in this study.

12 CC provides for the option to assign the isolate number for the work arbitrarily, if the chronological order is not easily ascertainable. In the case study, works can be ordered following either the tradition or general or specific reference works such as bibliographies, catalogues, thematic catalogues, and so on.

- date of each work associated with the author; and
- optionally, for a prospective Part 4 of the CC, the title of the works associated with the author.

Also, data elements correspondent to each facet must be available in a modular, shareable, and reusable form, to allow the large-scale production of class numbers for one or more national literatures.

Literary data in the Semantic Web

The first attempt to use a freely available and structured source of linked data relating to literary works and authors was based on Wikipedia, and especially on its Wikidata Query Service.[13] Data are accessible by a SPARQL end point, to be operated by SPARQL Language but also by a simplified user interface. In Wikidata, data are recorded as entities and properties, and some properties are useful to extract bibliographic data relating to authors and works. Anyway, the simplest approach is to start form a list of author-entities, reconcile them with Wikidata and extract further data by means of their specific properties (for example, *given name* (P735), *date of birth* (P569), *occupation* (P106), *notable work* (P800), etc.).[14]

The sample of 200 Italian authors extracted by means of pages available in Wikipedia[15] highlighted a few issues preventing a successful reuse of Wikidata to create automatic class numbers in CC:

1. the percentage of literary authors in the data retrieved by property occupation (P106) as author (Q2500638) was relatively low (15%);
2. values for property occupation (P106) are not recorded consistently, as the string values 'writer', 'author', 'poet' and 'novelist' are not associated to the 'creative person' in a foreseeable and coherent way; nevertheless, they are essential to identify any literary creator;
3. property 'author' (P50) could not be used, as it can apply to any kind of work (such as a scientific paper, for example), and not just to literary works; additionally, property 'creator' (P170) was too generic.

13 https://query.wikidata.org/

14 The author is extremely grateful to Andrea Zanni for his helpfulness and aid in these steps and for the useful discussion about data available in Wikidata and their quality.

15 https://it.wikisource.org/wiki/Categoria:Autori_del_XIX_secolo.

In general, the analysis of the sample data extracted from Wikidata showed interesting and useful data when relating to a single author, but also data lacking consistency, regularity and completeness when considered with respect to the universe of literary authors. Consequently, it was impossible to extract data relevant to all the authors and works of a specific national literature, or period, etc. The failed approach to extract coherent and complete data from Wikidata suggested finding a different, specialized source of linked open data; the prospective source had to contain authoritative data, created from the scratch and compliant with cataloguing standards. After an exploratory and quick analysis on both *data.bnf.fr*and*datos.bne.es*,[16] the data available from Bibliothèque nationale de France were chosen. Data necessary to produce class numbers for a sample of authors were extracted from *data.bnf.fr* by its SPARQL end point (http://data.bnf.fr/sparql/) (see Query in Appendix 1).

The sample

As the main goal of the case study was to evaluate the feasibility of the process, a restricted data sample was identified, collected and elaborated. Italian authors were chosen, according to the following requirements:

- authors born between 1800 and 1900;
- 'Italy' as country associated to the person;[17]
- 'Italian'" as language associated to the person;[18]
- 'Literature' (literal) as field of activity of the person.[19]

For each matched author, the following data were extracted from the SPARQL end point (Figure 1):

- URL of the work (column 1)
- Title of the work (column 2)
- Language of the work (column 3)
- URI of the date of the work (column 4)

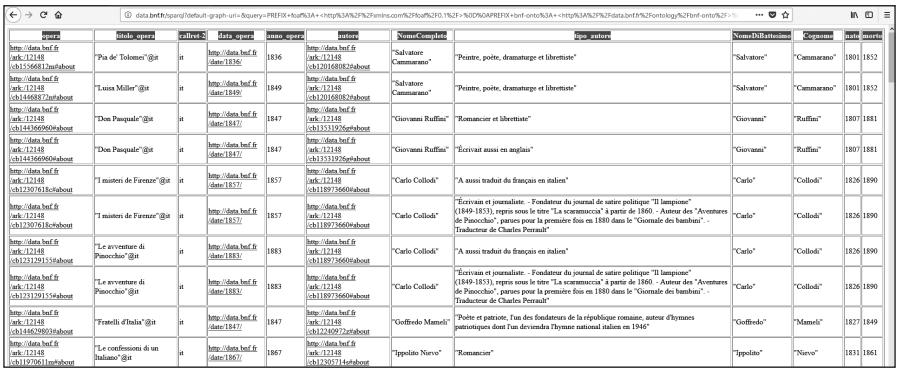
16 *datos.bne.es* showed a major issue: date of works was not available, while it was mandatory for the chronological arrangement of the works and their subsequent numbering. For this reason, *data.bnf.fr* was chosen as the source for the data sample.

17 Value http://id.loc.gov/vocabulary/countries/it for property "rdagroup2elements:countryAssociatedWithThePerson".

18 Valuehttp://id.loc.gov/vocabulary/iso639-2/ita for "rdagroup2elements:languageOfThePerson".

19 Value 'Littératures' for property "rdagroup2elements:fieldOfActivityOfThePerson".

- Date of the work (in numeric format) (column 5)
- Author URI (column 6)
- Full name of the author (in the form "Name Surname"; column 7)
- Field "Note" (usually containing a non-structured description of the author; (column 8)
- Given name of the author (column 9)
- Family name of the author (column 10)
- Birth date of the author (column 11)
- Death date of the author (column 12)



work	titolo_opera	caller1	data_opera	anno_opera	autore	NomeComplete	tipo_autore	NomeDiBattesimo	Cognome	nato	morto
http://data.bnf.fr/ark:/12148/cb15566812m#about	"Pia de' Tolomei"@it	it	http://data.bnf.fr/date/1836/	1836	http://data.bnf.fr/ark:/12148/cb12016808?#about	"Salvatore Cammarano"	"Peintre, poète, dramaturge et librettiste"	"Salvatore"	"Cammarano"	1801	1852
http://data.bnf.fr/ark:/12148/cb14468872m#about	"Luisa Miller"@it	it	http://data.bnf.fr/date/1849/	1849	http://data.bnf.fr/ark:/12148/cb12016808?#about	"Salvatore Cammarano"	"Peintre, poète, dramaturge et librettiste"	"Salvatore"	"Cammarano"	1801	1852
http://data.bnf.fr/ark:/12148/cb14456960t#about	"Don Pasquale"@it	it	http://data.bnf.fr/date/1847/	1847	http://data.bnf.fr/ark:/12148/cb13531926p#about	"Giovanni Ruffini"	"Romancier et librettiste"	"Giovanni"	"Ruffini"	1807	1881
http://data.bnf.fr/ark:/12148/cb14456960t#about	"Don Pasquale"@it	it	http://data.bnf.fr/date/1847/	1847	http://data.bnf.fr/ark:/12148/cb13531926p#about	"Giovanni Ruffini"	"Écrivait aussi en anglais"	"Giovanni"	"Ruffini"	1807	1881
http://data.bnf.fr/ark:/12148/cb12307618c#about	"I misteri de Firenze"@it	it	http://data.bnf.fr/date/1857/	1857	http://data.bnf.fr/ark:/12148/cb11897366f#about	"Carlo Collodi"	"A aussi traduit du français en italien"	"Carlo"	"Collodi"	1826	1890
http://data.bnf.fr/ark:/12148/cb12307618c#about	"I misteri de Firenze"@it	it	http://data.bnf.fr/date/1857/	1857	http://data.bnf.fr/ark:/12148/cb11897366f#about	"Carlo Collodi"	"Écrivain et journaliste. - Fondateur du journal de satire politique "Il lampione" (1849-1853), repris sous le titre "La scramuccia" à partir de 1860. - Auteur des "Aventures de Pinocchio", parues pour la première fois en 1880 dans le "Giornale des bambini". - Traducteur de Charles Perrault"	"Carlo"	"Collodi"	1826	1890
http://data.bnf.fr/ark:/12148/cb12312915t#about	"Le avventure di Pinocchio"@it	it	http://data.bnf.fr/date/1883/	1883	http://data.bnf.fr/ark:/12148/cb11897366f#about	"Carlo Collodi"	"A aussi traduit du français en italien"	"Carlo"	"Collodi"	1826	1890
http://data.bnf.fr/ark:/12148/cb12312915t#about	"Le avventure di Pinocchio"@it	it	http://data.bnf.fr/date/1883/	1883	http://data.bnf.fr/ark:/12148/cb11897366f#about	"Carlo Collodi"	"Écrivain et journaliste. - Fondateur du journal de satire politique "Il lampione" (1849-1853), repris sous le titre "La scramuccia" à partir de 1860. - Auteur des "Aventures de Pinocchio", parues pour la première fois en 1880 dans le "Giornale des bambini". - Traducteur de Charles Perrault"	"Carlo"	"Collodi"	1826	1890
http://data.bnf.fr/ark:/12148/cb14462980j#about	"Fratelli d'Italia"@it	it	http://data.bnf.fr/date/1847/	1847	http://data.bnf.fr/ark:/12148/cb12240972z#about	"Goffredo Mameli"	"Poète et patriote, l'un des fondateurs de la république romaine, auteur d'hymnes patriotiques dont l'un deviendra l'hymne national italien en 1946"	"Goffredo"	"Mameli"	1827	1849
http://data.bnf.fr/ark:/12148/cb11970611m#about	"Le confessioni di un Italiano"@it	it	http://data.bnf.fr/date/1867/	1867	http://data.bnf.fr/ark:/12148/cb12305714t#about	"Ippolito Nievo"	"Romancier"	"Ippolito"	"Nievo"	1831	1861

Figure 1 HTML output of the query on the *data.bnf.fr* SPARQL endpoint

Automatic collection of data was difficult with reference to one data element: data element 'field of activity', which was queried for 'Literature', was not available for an unquantifiable but relevant part of Italian authors recorded in *data.bnf.fr*. For this reason, the sample of Italian authors collected by the query resulted not to be complete. The data sample was of 31 authors and 97 literary works of the Italian Literature of the XIX and XX centuries. The total amount of authors and works included in the data sample are very far from the completeness; nevertheless, the sample was enough for the purpose of the case study.

Sample data processing

Sample data were exported in CSV format and imported in OpenRefine, a 'powerful tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.'[20] Data

20 http://openrefine.org/

processing required the following steps:

1) deduplication of records on the basis of identity of both Author's URI and URI of work;
2) arrangement by the author, date of the work and title of the work, in order to obtain the sequence required by CC work arrangement;
3) for each author, progressive numbering of his/her works arranged as per point 2;
4) clustering of authors for literary genre by data element Description (i.e. field 'Note' in *data.bnf.fr*).

Step no. 4 was the most difficult and longest one and required a small manual data processing by clustering function in OpenRefine. In fact, a relevant point to obtain correct CC class numbers in the main class *O Literature* was to identify each author's literary genre, as all the works by an author are classified by means of the author class number, which depends on the literary form associated with the author most. Information relating to the author's literary form were contained in the field "Note", but in an unstructured form (Figure 1; Column 8 'Tipo_Autore'). Nevertheless, the field 'Note' was useful even if unstructured, because it allowed to identify the prevailing literary form with the first characteristic used to describe the author: for instance, if the author's description was 'Romancier et auteur dramatique', Literary form facet in the CC class number was assumed to be "Romancier" as the right and relevant value and the author was associated to Novel as preferred literary form. To collocate author forms manually but more quickly, the value of the textual string containing the non-structured description was truncated to 20 characters. Just three authors (9%) required a specific manual classification, as their main literary role was unclear after the truncation.[21]

After cleaning data relating to literary form for each author, data were restructured automatically through the following steps:

columns were rearranged and selected to respect the order and the value required by facet formula;

verbalstring values were translated in their equivalent notation, by means of string transformation utilities available in OpenRefine, so that three new col-

21 In particular: Carlo Collodi was associated with the truncated string 'A aussi traduit du f', and Giovanni Ruffini was associated to the string 'Ecrivait aussi en an', both phrases without any reference to a literary form; Giorgio De Chirico was associated to the string 'Peintre, dessinateur' (a clear string but not usable to the literary classification). During the manual data revision Collodi and Ruffini were associated to Novel form, while De Chirico was excluded from the sample of Italian authors (even if he is associated to 843.91 in OCLC Classify; http://classify.oclc.org/classify2/).

umns (FaccettaForma, FaccettaAutore, and NumeroOpera) were automatically (see Figure 2, columns nos. 3, 5 and 7);

by rearrangement of authors' names (arranged alphabetically by family names), work title (arranged chronologically and, in suborder, alphabetically), and the sum of the values of the three newly created columns, a class number for each literary work was created in OpenRefine;

the CC class number tables for the complete sample of 31 Italian literary authors and 97 literary works were produced by the extraction of Full name of the author, + Work Title + Class number from OpenRefine to a text document (Appendix 1).

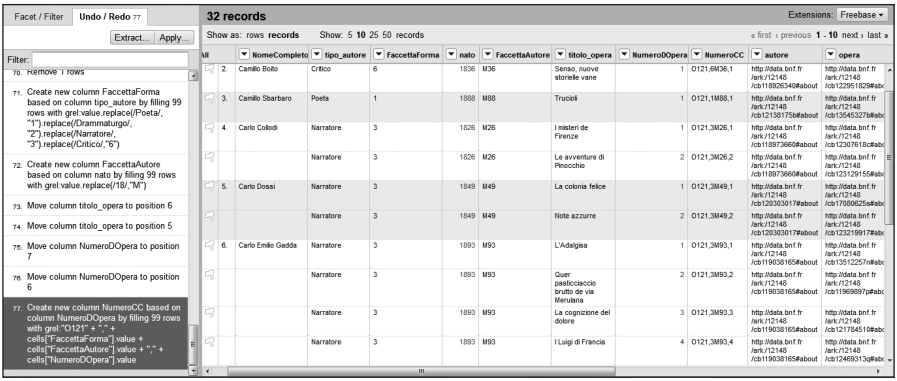


Figure 2 Data cleaning and transformation in OpenRefine

Discussion

The case study and its results put forward some interesting points with respect to the twofold starting issue, that is both whether an automatic, quick, affordable and large-scale creation of classification tables of CC class number for the literature is possible and, from a more general approach, which are the requirements for developing an automatic process for the creation of faceted class numbers. In the case study, the creation of class numbers for isolates of the main class *O Literature* relating to single works of specific authors resulted workable. Nevertheless, two possible issues emerged from the experimentation:

- during the first steps of the process, the extraction of linked open data on literary authors resulted partial, because only a part of the potentially relevant authors was associated with the tag 'literature'; and,

- during the step of identification of the relevant literary class for each author, data processing was not completely automated because the source data from *data.bnf.fr* were not structured (for example by means of a controlled vocabulary), but they were available only in a non-structured form.

Both issues can be bypassed in a relatively simple way, that is, widening the harvest of LOD to more than one namespace and reconciling data from many namespaces. These means should allow a more exact identification of literary authors and getting different attributes for the authors from various name spaces, for instance getting the author's name, surname, birth and death dates, and works from one source and author's preferred literary form and classification from another source. A specific feature of OpenRefine permits saving the whole data transformation process in a JSON file, that can be reused and applied to newly extracted data (for instance, a second sample of Italian literature authors, or the full extraction of French literature authors, and so on). The availability of a routine process assures replicability of the process to a larger scale, especially if controlled and high-quality data could be downloaded from many LOD sources.

As to the more general issue, the first thing to be observed is a prerequisite for the classification scheme. It must be based on regular semantics and syntaxes – that is to say on facets and facet formula, to permit class numbers to be obtained by an algorithm. This prerequisite prevents in a relevant way the applicability of the described process to classification schemes widely spread all over the word but not structured in such a way (e.g. Dewey Decimal Classification). An example of this is that the case study purposely took in exam the specific facet formula provided for single works of literary authors in the CC. In fact, Special Table in Part 3 of CC deals with these cases, and not with other kind of publications such as selections or collections of one or more authors. Moreover, the success of the process requires availability of semantic data with specific characteristics. It must be noted that:

- semantic data availability alone is not sufficient, as in the case of Wikidata, where property 'occupation' (P106) is provided for but data values are recorded inconsistently (i.e. lacking consistency and standardization provided for ICP (IFLA Section on Cataloguing Code, 2016)); so that, even if Wikidata is potentially usable as LOD, the way its data are produced and recorded did not assure sufficient consistence to their reuse in the case study;
- semantic data availability in one repository could not be sufficient alone, even if data are produced and recorded consistently; in order for data to

be reusable for the described process, their completeness with respect to the values of any facet in the facet formula is necessary. For example, at *datos.bne.es*, data relating to which and how many works were linked to an author were not accessible; and,

- complete and coherent semantic data availability is not sufficient alone, because their completeness must be checked also in relation to the their domain (for example, Literature); indeed, many Italian literature authors were not found by the query at *data.bnf.fr* SPARQL endpoint because of the lacking of value 'Littératures' in the property "field of activity of the person"; so that, accessible data allowed to perform the process, but results are likely to be quantitatively poor.

Therefore, for a more structured project aiming at large scale class number production a careful designed data harvest is mandatory in relation to the number and the quality and completeness of data elements.

Five lines of future research

The case study suggests a few hypotheses for further research. First, the issue of the difficult determination of the prevailing literary form associated with an author based on the unstructured biographical description obtained from *data.bnf.fr* could be fixed by the individuation of the third figure of DDC class number mainly associated with that author; for instance, an author mainly associated to DDC class number 853.xx could be associated to Italian novel. At least, till more reliable sources with this kind of data will be available.

Another possible line of research should include the individuation of other facet formulas within the CC scheme, and their analysis in order to determine the feasibility of such a process of automatic production of class numbers in other disciplinary fields. An example of such facet formulas in the CC are those relating to the classic works in other main classes, to be described by means of their relationships to other works (derived works, works based on works, and so on). Another example is that of main class *NR Music*, which is based on facets like space [P1], time [P2], a list of form of music publications [P3],[22] and instruments [M].[23]

22 For instance, Word/Libretto, Notation, Form, Keeping time, Dramatic music, Orchestral music.

23 For instance, *Wind instrument* (that includes *Pipe*, *Flute*, *Organ* etc.), *Stringed instrument* (that includes *Vina*, *Violin*, *Piano*, etc.).

A third interesting field of study could be the application of LOD to the production of book numbers, that is the part of classified call numbers that is used to distinguish among, and to identify, resources belonging in the same class. Ranganathan designed a facet formula for book numbers that is particularly interesting for many reasons: 1) it allows a unique identification of each book of a classified collection; 2) it is based on data elements usually available in MARC family records; 3) its application produces a sequence of resources that is fully compliant with a FRBR-zed arrangement of Expressions, Manifestations, and Items of a Work, by using core characteristics of each FRBR entity.[24] Since the identification of each item of a set of resources sharing the same class number is an issue not limited to CC but to any available scheme of classification, this line of research could be useful in a wide range of situations.

A fourth possible study could be directed to investigate and detect the presence of facet formulas in other classification schemes (for example, in class 400 Linguistics or in class 800 Literature of the Dewey Decimal Classification, or Universal Decimal Classification), to search further possible applications of the process.

The last, most complex and ambitious research goal is the design of a purely faceted classification scheme based mainly, if not exclusively, on data usually and easily available as linked open data. In many classification schemes, tables for common concepts such as language, space and time are provided. These concepts are largely and widely represented, recorded, and available as linked open data in the semantic web. They are a fundamental and relevant starting point for such a project.

References

Bianchini, C. (2012). Arrangement of FRBR Entities in Colon Classification Call Numbers. *Cataloging & Classification Quarterly*, 50(5-7), 1-21, https://doi.org/10.1080/01639374.2012.679877.

Bianchini, C. (2017). *Book number: uno strumento per l'organizzazione delle collezioni. Manuale ad uso dei bibliotecari*. Milano: Editrice Bibliografica.

IFLA Section on Cataloguing Code. (2016). *Statement of International Catalogu-*

24 Core characteristics for the book number facet formula are *language* (referred to as the Expression) and *date of publication* (relating to the Manifestation). For further readings on the book number and its relationships with FRBR, see (Bianchini, 2012), and, more recently (in Italian), (Bianchini, 2017, pp. 73-94).

ing Principles (ICP). A. Galeffi, M.V. Bertolini, R.L. Bothmann, E. Escolano Rodríguez and D. McGarry (eds.). Den Haag: IFLA.

Ranganathan, S.R. (1959). *Elements of library classification : Based on lectures delivered at the University of Bombay ... 1944, and in the schools of librarianship in Great Britain ... 1956*. B. I. Palmer (ed.) (2nd ed.). London: The Association of Assistant Librarians.

Ranganathan, S.R. (1963). *Colon classification. Basic classification* (6th ed. rev). Bangalore: Sarada Ranganathan Endowment for Library Science.

Ranganathan, S.R. (1964). *Classified catalogue code with additional rules for Dictionary catalogue code* (5th ed.). Bombay: Asia Publishing House.

Ranganathan, S.R. (1965). *A descriptive account of the Colon Classification*. Bangalore: Sarada Ranganathan Endowment for Library Science.

Ranganathan, S.R. (1967). *Prolegomena to Library Classification* (3rd ed.). Bombay: Asia Publishing House.

Tartaglia, S. (1998). *Ordine di citazione e principio di faccettazione nella Classificazione Decimale Dewey*. Udine: Forum.

Varghese, M. (2010). Relevance of a Classified Catalogue in the FRBR perspective and a proposed Model with ISBD descriptions and Faceted Class number as Key attribute. *Cataloging & Classification Quarterly*, 46(3), 281-304.

Appendix no.1

Special Tables. Italian Literature

This Appendix contains CC class numbers of Italian literature works obtained processing linked open data extracted from *data.bnf.fr*.

Titles of the works just indicate the work; they neither correspond to the titles on the first edition nor represent a 'uniform title'; in fact, identification is to be obtained by CC work number. As all the works are classics, form of their titles varies depending on the reference source (for instance, *Enciclopedia Treccani*, BNCF opac, Wikipedia). Therefore, the form of the title is taken from *Dizionario letterario Bompiani delle opere e dei personaggi di tutti i tempi e di tutte le letterature* (Milano, Bompiani, 1964-1972), when available. Otherwise, it is taken directly from *data.bnf.fr*.

Author	Work	Work class number
Corrado Alvaro	*L'uomo è forte*	O121,3M95,1
	L'età breve	O121,3M95,2
	Memorie del mondo sommerso	O121,3M95,3
	Lunga notte di Medea	O121,3M95,4
Luigi Antonelli	*La donna in vetrina*	O121,2M77,1
Camillo Boito	*Senso*	O121,6M36,1
Anna Bonacci	*L'ora della fantasia*	O121,2M92,1
Salvatore Cammarano	*Pia de' Tolomei*	O121,2M01,1
	Luisa Miller	O121,2M01,2
Dino Campana	*Canti orfici*	O121,1M85,1
Carlo Collodi	*I misteri di Firenze*	O121,3M26,1
	Le avventure di Pinocchio	O121,3M26,2
Edmondo De Amicis	*La vita militare : bozzetti*	O121,1M46,1
	Cuore	O121,1M46,2
	Primo maggio	O121,1M46,3
Grazia Deledda	*Canne al vento*	O121,3M71,1
	Il segreto dell'uomo solitario	O121,3M71,2
Emilio De Marchi	*Demetrio Pianelli*	O121,3M51,1
Federico De Roberto	*La sorte*	O121,3M61,1
	I viceré	O121,3M61,2
Carlo Dossi	*La colonia felice*	O121,3M49,1
	Note azzurre	O121,3M49,2
Carlo Emilio Gadda	*L'Adalgisa*	O121,3M93,1
	Quer pasticciaccio brutto de via Merulana	O121,3M93,2
	La cognizione del dolore	O121,3M93,3
	I Luigi di Francia	O121,3M93,4
	La meccanica	O121,3M93,5
	Meditazione milanese	O121,3M93,6
Corrado Govoni	*Le fiale*	O121,1M84,1
Guido Gozzano	*La via del rifugio*	O121,1M83,1
	I colloqui	O121,1M83,2

Author	Work	Work class number
Luigi Illica	*Andrea Chénier*	O121,2M57,1
Goffredo Mameli	*Fratelli d'Italia*	O121,1M27,1
FilippoTommaso Marinetti	*Mafarka, il futurista*	O121,1M76,1
	Parole in libertà	O121,1M76,2
	Gli indomabili	O121,1M76,3
	Teatro della sorpresa	O121,1M76,4
	Cinque sintesi radiofoniche	O121,1M76,5
Eugenio Montale	*Ossi di seppia*	O121,1M96,1
	Il balcone dentro alle occasioni	O121,1M96,2
	Le occasioni	O121,1M96,3
	Mediterraneo	O121,1M96,4
	Mottetti	O121,1M96,5
	Finisterre	O121,1M96,6
	Farfalla di Dinard	O121,1M96,7
	L'anguilla	O121,1M96,8
	La bufera e altro	O121,1M96,9
	Satura	O121,1M96,10
	Diario postumo	O121,1M96,11
Dario Niccodemi	*Prete Pero*	O121,2M74,1
Ippolito Nievo	*Confessioni di un Italiano*	O121,3M31,1
Luigi Pirandello	*L'esclusa*	O121,2M67,1
	Il turno	O121,2M67,2
	Il fu Mattia Pascal	O121,2M67,3
	L'umorismo	O121,2M67,4
	I vecchi e i giovani	O121,2M67,5
	Quaderni di Serafino Gubbio operatore	O121,2M67,6
	La giara	O121,2M67,7
	Il giuoco delle parti	O121,2M67,8
	Sei personaggi in cerca d'autore	O121,2M67,9
	Enrico IV	O121,2M67,10
	Novelle per un anno	O121,2M67,11
	Ciascuno a suo modo	O121,2M67,12
	Uno, nessuno e centomila	O121,2M67,13

Author	Work	Work class number
	Lazzaro	O121,2M67,14
	Questa sera si recita a soggetto	O121,2M67,15
	I giganti della montagna	O121,2M67,16
Giovanni Ruffini	*Don Pasquale*	O121,3M07,1
Umberto Saba	*Canzoniere*	O121,1M83,1
	Autobiografia	O121,1M83,2
	Scorciatoie e raccontini	O121,1M83,3
Camillo Sbarbaro	*Trucioli*	O121,1M88,1
Renato Serra	*Esame di coscienza di un letterato*	O121,6M84,1
Scipio Slataper	*Il mio carso*	O121,3M88,1
Italo Svevo	*Una vita*	O121,3M61,1
	Senilità	O121,3M61,2
	La coscienza di Zeno	O121,3M61,3
	Una burla riuscita	O121,3M61,4
	La novella del buon vecchio e della bella fanciulla	O121,3M61,5
Giuseppe Tomasi di Lampedusa	*Il gattopardo*	O121,3M96,1
	La sirena[25]	O121,3M96,2
Giuseppe Ungaretti	*Il porto sepolto*	O121,1M88,1
	Sentimento del tempo	O121,1M88,2
	Allegria	O121,1M88,3
Giovanni Verga	*I Carbonari della montagna*	O121,3M40,1
	Sulle lagune	O121,3M40,2
	Storia di una capinera	O121,3M40,3
	Rosso Malpelo	O121,3M40,4
	Vita dei campi	O121,3M40,5
	I Malavoglia	O121,3M40,6
	Novelle rusticane	O121,3M40,7
	Cavalleria rusticana	O121,3M40,8
	Mastro don Gesualdo	O121,3M40,9
	Don Candeloro e C.i	O121,3M40,10
	La caccia alla volpe	O121,3M40,11
	Dal tuo al mio	O121,3M40,12

OD SEMANTIČKOG *WEBA* DO FACETNE KLASIFIKACIJE
STUDIJA SLUČAJA I PET PRAVACA BUDUĆEG ISTRAŽIVANJA

Ključne riječi:
facetna klasifikacija, semantički web, Klasifikacija s dvotočkom, ***data.bnf.fr****, klasifikacija talijanske književnosti*

Sažetak

Opisana je studija slučaja automatskog stvaranja brojeva facetne klasifikacije za posebne teme, koje polazi od podataka dostupnih na semantičkom webu. Cilj je istražiti dva pitanja: 1) Je li moguće automatski, brzo, ekonomično i u velikom opsegu generirati brojeve Klasifikacije s dvotočkom za književne klasike? i 2) Općenito: Koji su uvjeti za automatsko generiranje brojeva facetne klasifikacijske sheme ? Kakva su istraživanja potrebna za razvoj takvog procesa? Studija slučaja usredotočena je na podatke izvučene iz data.bnf.fr i na njihovu ponovnu uporabu u cilju dobivanja točnih i potpunih brojeva Klasifikacije s dvotočkom za uzorak talijanskih književnika (rođenih od 1800. do 1900.). Predstavljen je postupak kreiranja klasifikacijskih brojeva koji polazi od identifikacije i ekstrakcije podataka te njihovog čišćenja, transformacije i prijevoda u klasifikacijsku notaciju pomoću alata Open Refine. Raspravlja se o rezultatima studije slučaja i njihovoj važnosti za početna pitanja te se identificiraju istraživačka pitanja za moguće buduće pravce razvoja.