DESIGNING A DATABASE TO ASSIST LEGAL THINKING: A NEW APPROACH TO INDEXING USING FACETS

Günter Reiner / Michelle Cumyn / Michèle Hudon / Sabine Mas

Professor of Law, Fakultät für Wirtschafts- und Sozialwissenschaften, Universität der Bundeswehr Hamburg
Holstenhofweg 85, D-22043 Hamburg, Germany, DE
guenter.reiner@unibwh.de; http://hsu-hh.de/reiner

Professor, Faculté de droit, Université Laval, Pavillon Charles-De Koninck
1030, av. des Sciences-Humaines, Québec (Québec) G1V 0A6, Canada, CA
michelle.cumyn@fd.ulaval.ca; https://www.fd.ulaval.ca/faculte/professeurs/michelle-cumyn

Professor, École de bibliothéconomie et des sciences de l'information, Université de Montréal
Pavillon Lionel-Groulx, 3150, rue Jean-Brillant, Montréal (Québec) H3T 1N8, Canada, CA
sabine.mas@umontreal.ca; http://mas.ebsi.umontreal.ca/

Associate Professor, École de bibliothéconomie et des sciences de l'information, Université de Montréal
Pavillon Lionel-Groulx, 3150, rue Jean-Brillant, Montréal (Québec) H3T 1N8, Canada, CA
michele.hudon@umontreal.ca; https://www.ebsi.umontreal.ca/repertoire-ecole/vue/hudon-michele/

Keywords: ***Automatic indexing, controlled vocabulary, database, facet, faceted classification, indexing, legal information retrieval, legal reasoning, ontology, structured indexing***

Abstract: *A well-designed database supports legal thinking by bringing together the request for information and the information contained within the documents of the database. Indexing is one of the oldest means to achieve this goal, and yet there is still a lot of potential for improvement. What is missing is a syntactic structuring of semantic indexing. This is the subject of the interdisciplinary project described below, which is currently entering the testing phase. Inspired by the theory of faceted classification, we have built a prototype database containing 2,500 cases in the areas of administrative law, labour law and the law of obligations. Using a controlled vocabulary, we have (manually) indexed such decisions with keywords, assigning each one to one of six predefined categories or facets (Person, Action, Thing, Context, Legal category and Sanction). In this way, we have formalized the legal essence of the database content in a way that is not unlike an ontology. Our hypothesis is that our model will improve search results, facilitating the search for cases with similar facts and supporting legal thinking by revealing connections between facts and legal consequences that co-occur within decisions of the database. A further step will be to consider whether our scheme may serve as a tool for the automatic indexing of court decisions.*

1. Introduction

Thinking (here synonymous with reasoning) involves the processing of information. Legal thinking involves the processing of legal and legally relevant information. In this respect, the computer, *i.e.* the machine that specializes in data processing, seems predestined for use in law. Law, *i.e.* information that can be found in statutes, cases and legal literature, consists of rules; and legal thinking claims to apply rules. However, legal thinking is not merely deductive and legal norms are not rules in the strict, logical sense of the word. Although a legal argument must be rational to be accepted, it does not fully comply with formal logic (Cumyn 2015, at 72–76). Legal rules rely on linguistic forms and are characterised by uncertainty, as with all linguistic communication. Moreover, many matters are not resolved by a given rule, but the legislator leaves them to the judgment of the courts, to be decided on the basis of deliberately indeterminate legal terms, broad principles or precedents. Finally, the methods for interpreting and applying legal provisions and for deriving rules from precedents also lack precision and predictability.

The non-deductive side of applying the law is characterised by discretion, *i.e.* the weighing of possible outcomes in reaching a decision. Even if it were possible to design an algorithm for that purpose (*e.g.* Alexy 2003, at 443–448, and his «weight formula»), the computer could never replace a human decider, because in order to persuade others, a legal decision, in addition to rationality, must show a measure of empathy, of which machines are not capable (Reiner 2019, at II.3.c). Yet it is conceivable that computers, thanks to their ability to recognize patterns and to learn by themselves, will be able to help decision-makers by processing legal documents, such as cases and scholarly works, with a view to suggesting possible arguments and solutions, as well as identifying untenable ones as such.

The potential for preparing or supporting legal decision-making through the use of computers is obvious. In fact, the automated application of the law has already crossed the threshold into reality, despite legitimate concerns about its ethical implications (Rouvroy 2018, at IV, «Un séisme épistémique»). For instance, the German General Tax Act (*Abgabenordnung*) permits revenue authorities «to use fully automated processes to conduct, correct, withdraw, revoke, cancel or amend tax assessments» based on the information at their disposal and the data supplied by the taxpayer; this process is actually not a fully automated subsumption because it is the taxpayer himself who characterizes his data by selecting the appropriate headings in the standardized electronic form. In Canada, the federal government assesses immigration files using a predictive algorithm (Ling 2018). These examples illustrate the range of automation initiatives, from a formal, rule-based approach for typical («routine») cases with clearly structured data to one that facilitates the exercise of discretion through pattern recognition.

There already exists extensive, hardly manageable theoretical and practical preliminary work on these topics. Research initiatives reached an initial peak in the 1970s (e.g. Kilian 1974) and then subsided somewhat, but in recent years, have picked up again. At first, attempts were made to formalise the application of legal rules, and to reproduce it through computer programming, using a rule-based approach. More recently, the preferred approach has been to replicate or assist the non-deductive weighing of possible outcomes through pattern recognition. Such research, which in some cases has been carried out with considerable effort (*e.g.* the empirical research by Gerathewohl 1987), has not yet fulfilled its promise, and its current applications are limited in scope.

Computer support is undoubtedly most advanced in the area of *legal databases*, where the challenge is to identify and procure relevant legal documents that match a query. This is not to be underestimated, but there is still much room for improvement. As we will show, the research interface in a legal database is, or can become, much more than just a technical aid. It is our belief that indexing legal documents using facets would make valuable information available to the user, enhancing the performance of the database; that it would also pave the way for automated indexing; and finally, that it may provide a useful conceptual structure for the development of AI initiatives in the legal domain. After providing some thoughts on the ways in which databases might better support legal thinking, we present our indexing model and prototype and consider avenues for automated indexing.

2. Requirements for a database that supports legal thinking

2.1. The task of legal research

Legal databases in their current form increasingly contribute *directly* to legal thinking. Legal research in databases and legal thinking are closely interwoven. It is no coincidence if the term «research» has a dual meaning:

1) Searching for information in a library or database
2) Scientific research, *i.e.* exploration and reflection

Thinking requires knowledge before it can generate knowledge (Krathwohl 2002, at 212, 214, referring to the «knowledge dimension» of the revised Bloom's Taxonomy of educational objectives). This also applies to the search for information, which is necessarily preceded by an act of thought. In this respect, searching for information is almost a paradox. If one is looking for information contained in a document, one must have a certain idea of the information one is looking for in order to search for documents that contain it (*cf.* Kuhlthau 1991, at 362). Searching is a means of thinking, and conversely thinking is a means of searching. If it is possible to recognize the conceptual map of a certain domain and to incorporate it into the structure of the database, search results may be optimized. A *faceted classification* can help to accomplish this task: it creates a uniform grid for asking questions of an unknown document and the information it contains (*e.g.* «Which persons, which actions, which things were involved?»).

In the *legal* field, the two meanings of legal research, search for information and exploration/reflection, have in common that they aim to accomplish a legal task (in a broad sense). Such a legal task may simply consist of searching for a specific document or set of documents that match predefined criteria. For example, one wishes to find a case that the Court of Appeal handed down on April 1st 2010 or one is interested in cases awarding compensation for pain and suffering due to a concussion. In these examples, the decisive search criteria are already included in the task description; therefore no legal thinking is involved.

Other searches truly require and generate legal knowledge, because they aim at answering a *legal* question, *i.e.* a question whose answer requires the application of the law, by means of information contained in legal documents. Most frequently the user is looking for the legal characterization of certain facts, in order to ascertain the legal consequences attached to them. In other situations, the search is for possible facts that would generate a desired legal consequence, for example how a disclaimer should be phrased in order to be effective. Faced with a legal issue, the person looking for an answer tries to characterise the issue and/or the (real or anticipated) facts underlying the issue in order to determine what elements are missing towards a solution. The more experienced the user, the more their search will be targeted to the relevant legal categories and concepts representing the authoritative rules.

Such legal research requires an initial impulse that presupposes structural and conceptual knowledge, *i.e.* the building of a mental model of the topical area (Komlodi 2002). The information seeker needs to break down the question into search terms (unless the database is capable of interpreting a question using natural language, which until now works only for very simple questions). If legal concepts are used as search terms, the need for structural analysis goes without saying, but even if users confine themselves to searching with terms describing certain facts, they should at least know or have a feeling for what part of reality is likely to be relevant. (Another possibility, the establishment of which is very time-consuming for the database provider, would be an interface offering a linked decision tree based on a number of typical questions.)

Structural knowledge gained by the user can itself be the result of a search. Quite often, the results of one's initial query help to reframe it with the use of different search terms. This can be done either directly by looking at the results list, or indirectly by reading a few documents selected from the results list. Failures are in themselves instructive («multi-directionality in information behaviour», Godbold 2006), yet they cost time and effort. The challenge is to guide users as quickly and as directly as possible to the required information and to the relevant documents by helping them select the right combination of search terms. That should be the aim of computer scientists, information scientists and lawyers working to improve the performance of legal databases.

2.2. The assistance provided by legal databases

The legal database should assist the user who needs to apply the law in order to solve a legal question or problem. This assistance is mainly provided *indirectly* through the documents contained in the database and delivered to the user. It requires matching the information request of the user with the information contained within documents of the database.

There are two sides from which matching (and legal thinking) can be promoted, from the user's side and from the database provider's side.

1) Sitting in front of the search interface, the user tries to design an effective query. In order to do so, they anticipate the possible contents of the documents they are looking for, using distinctive wording (legal concepts; relevant elements of fact) and extended (Boolean) search functions. As J. Farradane pointed out, Boolean logic is «inadequate or misleading for structuring the terms in a question» and «expresses accurately only a small part of the relations between [search] terms» with which a user tries to describe their thinking (Farradane 1979, at 267).
2) Behind the search interface, the database provider prepares (manually or automatically) the results (documents) the user is looking for in such a way that they approach as closely as possible the search terms that the user is expected to choose.

An effective database is one where the mutual anticipation of users and provider converge. The use of an index is the traditional method to achieve this goal (cf. Prestel 1971, at 51 *et seq*, comparing empirically manual keyword indexing with a full-text system). Commercial databases continue to provide manual semantic indexing using keywords (in addition to full-text indexing, which improves the performance of a full-text search). The costs involved, and the doubts sometimes expressed concerning the utility of an index, have not (yet) discouraged their use (see Gross *et al.* 2015).

By revealing the indexed terms associated with documents of the results list and by showing in short excerpts their textual context, the interface itself (including the results list) generates potentially useful information in relation to the user's legal question by showing relationships:

- between search terms in a full-text search;
- between indexed terms (keywords);
- between facts and legal consequences;
- between documents (see the function «searching for similar texts» in the German database *juris*).

2.3. The need to improve semantic indexing

Despite a long tradition of legal indexing, which goes back centuries, there is still much potential for improvement. Both formal characteristics relating to the type and origin of a document, and substantive information regarding the content or subject-matter of a document, may be represented by index terms. Formal indexing is well-developed in the context of legal databases (*e.g.* deciding authority, jurisdiction, date, file number). One cannot say the same of semantic indexing, which in our view is insufficient. In describing the essential contents of a document, indexed terms create an additional semantic layer, which can be legal, factual or both, thus revealing the relationship between law and reality; but so far, this potential has not been fully exploited. Usually, semantic indexing is either not structured at all – the indexer chooses whatever elements they consider important – or it must follow a predefined sequence flowing from the relevant field (subject area) of law. A well-known example is Westlaw's «Key Number System», which divides the legal order, in the manner of a textbook or manual with multiple levels, into individual areas and aspects to which it assigns approximately 100.000 searchable numbers.

What is missing is a flexible syntactic-functional structuring («syntactization») of semantic indexing (Schweighofer / Lachmeyer 2017, at 2). Our research program investigates the use of facets to achieve that aim. It is hoped that semantic indexing using facets will enable users to design targeted searches and will present more clearly the information that the interface itself generates. The co-occurrence of indexed terms within the different documents of the database is not random: it represents meaningful connections between concepts. Every case may be indexed using a set of terms that grasp its core content, *i.e.* content that is legally significant, in such a way that it follows a train of legal thought albeit in simple, reduced form.

3. A new approach to indexing using facets

3.1. Presentation of the approach

Inspired by the theory of faceted classification, we have developed an indexing model that follows a content grid of six predefined categories (or facets in a broad sense) (Cumyn *et al.* 2018). This structure represents the «grammar» of legal information contained in documents of the database. Our hypothesis is that this model will facilitate the search for cases with similar facts (as regards the relevant law) and support legal thinking by revealing connections between facts, legal categories and sanctions that co-occur within decisions of the database.

In order to test our approach, we have built a Web application prototype with Python programming language and the Django framework. The application, that serves both as an indexing tool and database, contains 2,500 cases from Québec (written in French for the most part) in the areas of administrative law, labour law and the law of obligations. Using a controlled and structured vocabulary (thesaurus, c.f. ISO 25964-1:2011 (E), at 2.62) developed (in French) incrementally under the supervision of an expert librarian, we have manually indexed each case. The thesaurus contains scope notes and basic semantic relations: broader term, narrower term, preferred term, non-preferred term (for synonyms) and related term. At the time of indexing, keywords (drawn from the thesaurus) are assigned to the appropriate facets in the case content description. The indexing tool provides a basic search interface so that indexers can visualize the indexing of prior cases. Indexing has been completed, and our model is now entering the testing phase (for details see Cumyn et al 2018, at 885–886).

Faceted classification and indexing schemes are not new, even in law. The method and the theory of facet analysis were conceived by the Indian mathematician and librarian S. R. Ranganathan (1892–1972), inspired by the emergence of specialized, micro, and interdisciplinary subjects, with which existing classifications were unable to cope (Satija 2017, at 292). Developed in the 1920s, Ranganathan's theory formed the basis for the Colon Classification, published in seven editions from 1933 to 1987 (see Satija 2017, at 291). According to Ranganathan, every subject matter or unit of knowledge can be broken down according to five fundamental, broad and mutually exclusive categories, which he originally called «train of characteristics» and later «facets» and which he arranged pursuant to their «decreasing concreteness» as follows (the so-called PMEST formula): 1) *Personality* (entities or things), 2) *Matter* (materials or constituents of things), 3) *Energy* (actions or activities), 4) *Space*, and 5) *Time* (see the «Postulates» 1 and 2 of Facet Analysis, Ranganathan 1959, at 67 et seq, 37).

Ranganathan's facets, which one can think of as aspects or dimensions of a topic, revolutionized the field of knowledge organization thanks to the Classification Research Group (CRG) founded in the United Kingdom in 1952. As explained by V. Broughton , who later became a member of the group herself, «[t]he real genius of the CRG lies in its shaping of facet analysis as a generalised methodology for information retrieval» (Broughton 2011, «Conclusion »). In 1955, the CRG published an interim report stating that «[t]he terms subsumed under a given genus [...] are not all derived from that·genus by differentiation using a single characteristic of division», but that they «can be sorted into groups or facets, each of which is derived from that genus by a different characteristic». They concluded that «[i]n order adequately to display the linkage of related terms it is necessary to have faceted classification schedules» (CRG 1955, at 2) and without admitting the existence of five fundamental categories they recognized «that the use of a provisional set of categories - formulated by practical examination of certain subject fields – is helpful in making a first approach to other subjects» (at 9). The 2nd edition (1977) of the Bliss Bibliographic Classification is the example of a «universal system of bibliographic classification built on the facet analytical principles developed by the CRG»

(Broughton 2001, at 74). Ranganathan's original facets were there expanded to 13 categories representing a «production process», and being «particularly suitable for the analysis and organization of terms in technology»: Thing/entity, Kind, Part, Property, Material, Process, Operation, Patient, Product, By-product, Agent, Space and Time (Broughton 2001, at 79; Cumyn et al. 2018, at 884).

The universalist ambition of faceted classification suggests that it should be extended to the legal domain, but so far only a few attempts were made. After participating in the CRG, Broughton sought to implement the Bliss Classification in the legal documentation (Broughton 2010, at 37–40). Her goal was to design a new list of subject headings such as one finds in library catalogues, using the CRG's facets. Assuming that «[i]n the context of any given subject field, usually only a smaller number of the categories are relevant» (Broughton 2010, at 37), she defined the following «five significant» facets in law: *Jurisdictions (Space)*, *Substantive law (Thing; Personality)*, *Legal practice and procedure (Process and Operation; Energy)*, *Attributes and principles of law*, *Jurisprudence (Property; Matter)*, and a residual category named *Common subdivisions* that includes *Agents* and *Time*. As we have indicated in brackets, each one of these facets can be traced to the CRG's and Ranganathan's schemes. We are not aware that this system has been implemented in practice.

In a similar vein, Sweet & Maxwell, a British legal publisher, has created a «Legal Taxonomy» for structuring its indexers' thesaurus in accordance with the principles of facet analysis (Scott / Smith 2010, at 217). The terms to be classified were drawn from an index designed for legal periodicals. The scheme is intended to be used for all the publisher's printed and online publications relating to English law, including WestlawUK. The new classification has been successfully implemented. The taxonomy, which is updated regularly, is available online (http://2.sweetandmaxwell.co.uk/online/taxonomy/). It is unusual in that top-level terms designate 111 hierarchized legal subject areas ranging from «accountancy» to «water law», which it is said should not stray «too far from some of the ‹natural› divisions of subject matter» (Scott / Smith 2010, at 219). Facets are introduced within each subject area, with a view to making its internal structure more consistent and predictable for indexers and users alike. Twenty «standard facets» are identified for that purpose: *Attributes, Courts, Civil procedure rules, Documents, Entities, Events and actions, Judgments and orders, Liabilities, Markets, Notices and orders, Payments, Persons, Place, Policies, Powers rights and duties, Principles, Statements, Time, Tribunals, Vitiating factors* (Scott / Smith 2010, at 219).

The manner in which facet analysis was applied to the legal domain by Broughton and by Sweet & Maxwell does not fulfil the promise of faceted schemes; nor do they meet our requirements, for several reasons. The hierarchical classification of subject headings proposed by Broughton is rigid, whereas faceted schemes are supposed to be flexible (Kwasnik 1999, at 39–41). Sweet & Maxwell provide flexibility by allowing the same terms to be repeated within each subject area, and sometimes under more than one facet, but the resulting taxonomy is over-developed and redundant, whereas a classification scheme should be economical. Moreover, Sweet & Maxwell's use of subject areas as top terms encourages users to limit their search to a given subject area, whereas facets ought to promote different perspectives or points of entry (Hudon / el Hadi 2010, at 24). Many legal questions cut across subject areas, and a legal database, especially one that embraces facet analysis, should engage jurists to think outside the silo of a given practice area or specialisation.

In addition, Broughton and Sweet & Maxwell worked with existing lists of subject headings and index terms, and applied facet analysis to the underlying concepts without addressing their nature or function. Concepts used by lawyers are often ambiguous in that they relate simultaneously or alternatively to a set of facts on the one hand or to a legal category, defined as a set of rules, on the other (Cumyn / Gosselin 2016, at 335–338). For example, there may have been a theft: this is a description of the facts. Different legal consequences may follow: one might apply the criminal law relating to theft, or one might turn to labour law, property law, the law of insurance, etc., depending on the legal question one has to answer in relation to the facts. We discovered that applying facet analysis to the names of legal categories is both difficult and confusing. However, we also found that applying facet analysis to the facts of each case is not only feasible, it appears to highlight relevant information about the case. A related point is that traditional legal indexes and classifications favour the use of abstract legal concepts over those that merely describe the facts. Thus we came to develop our own indexing scheme based on facet analysis of the cases themselves and consequently, we created our own controlled vocabulary, instead of relying on existing thesauri. To clarify the use of concepts, which, like theft, may represent either the facts of a case or applicable rules of law, we created separate categories (or facets in a broad sense) for dealing with the factual elements of a case and the legal consequences that follow. This also reflects the structure of legal thinking outlined above: answering a legal question frequently requires the characterisation of certain facts, i.e. identifying the legal categories that may apply to such facts; or in reverse, knowing the facts that would trigger the application of a given legal category.

Finally, facets are commonly used as a browsing tool, and this is arguably their most attractive feature, if one is designing the search interface of a database. Faceted search interfaces allow the user to filter results by selecting from the set of indexed terms (*labels*, *cf.* Hearst 2006, at 26) associated with each facet. To be effective, facets must be intuitive and limited in number.

We found Ranganathan's PMEST formula to be the simplest and the most intuitive. We noted its similarity with Gaius' famous tripartite division, well-known to all Western legal systems. Gaius declared that «all our law is about persons, things or actions» (*Institutiones*, 1.8, http://www.thelatinlibrary.com/gaius1.html#8; also see Birks 1997, at 5). For Ranganathan, persons and things belong to the facet Personality, but in law, the distinction is so fundamental that it is necessary to provide separate facets. After experimenting with various schemes, we arrived at the following six facets:

- *Person*: a natural or legal person, body or entity that has decision-making authority (*e.g.* lawyer, farm worker, Ministry of Justice, committee).
- *Action*: an act, activity or decision that is governed by law or that has legal consequences (*e.g.* accident, sale, dismissal).
- *Thing*: a tangible or intangible, concrete or abstract entity that is subject to a legal framework or protected by law; typically, the object of litigation or the instrument of an action (*e.g.* car, deed, licence).
- *Context*: an additional element that is essential for the treatment of a legal problem with respect to time and place or regarding the character of a thing or action, including a cause or consequence (*e.g.* delay, age, deficiency).

The above are facets in the narrow sense, since they are derived from Ranganathan's scheme. Like his fundamental facets, they are intended to grasp humans' perception of reality. The facets capture the factual basis of a case in its legally relevant dimensions, since the law regulates the conduct of persons with respect to things. The systematic and structured indexing of facts and acts that form the basis for legal judgments and the ability to mark them as facts is a novelty in legal databases (for a similar proposal by a German author see Moeller 1993, at 184, who suggests that the search interface of the German database *juris* should be divided into two fields, one for «legal searches» and one for «factual searches»). In this way, our system responds to the needs of users in the computer age who, thanks to the full text search, are inclined to look for cases using facts rather than legal concepts (Strauch 2017, at 605–609; see also Cumyn *et al.* 2018, at 883).

The last two facets (in a broad sense) are devoted to legal consequences:

- *Legal category*: a concept referring to a set of rules or precedents, typically found in a statute or part of a statute, a leading case or a line of case law (*e.g.* assignment of claim, right to counsel).
- *Sanction*: remedy, compensation, punishment or other form of relief resulting from the application of legal rules (*e.g.* punitive damages, interlocutory injunction).

The latter are facets only in a broad sense, because they belong entirely to the legal domain and do not reflect any of the facets envisioned by Ranganathan or the CRG. They resemble the classifications of legal concepts which one finds in conventional databases, or the subject areas of Sweet & Maxwell. However, they are more homogenous and less ambiguous, since they do not refer to the facts of a case, but describe only the legal consequences attached to the facts, which are represented by the above-mentioned facets in the narrow sense. Our decision to distinguish legal categories and sanctions is based on the special importance of the latter. Legal questions are often questions about applicable – solicited or feared – sanctions. Then it may be advisable to start by searching for legal rules that impose those sanctions.

3.2. The grammar analogy: a syntax of legal information

Our facets mediate between the experts in knowledge organization and lawyers' own understanding of their discipline (Cumyn *et al.* 2018, at 885). They bring together Gaius' famous tripartite division and Ranganathan's PMEST formula (see the comparative table, *id.*). The scheme is so simple, that it might even seem trivial, yet it is hoped that this will make it intuitive and easy to grasp for users. It is interesting that facet analysis finds a parallel in linguistics, especially in grammar (*cf.* Maniez 1999, at 252–253), broadly defined as the whole system of rules or procedures that make up a given language. This will be explained using Ranganathan's scheme.

The PMEST facets correspond to four dimensions or functions on which numerous languages are built. They are represented by the word types «noun», «verb», «adjective», and «adverb»:

- Personality ≈ noun
- Energy ≈ verb
- Matter ≈ adjective
- Time ≈ adverb
- Space ≈ adverb.

Surely this comparison should not be taken literally. Index terms assigned to all five facets are nouns, in accordance with accepted standards regarding the construction of thesauri (see ISO 25964-1-2011, at 6.3.1). The distinction between the four word types seems to be based on a similar perception of reality as the one underlying the distinction between facets P, M, E and T/S. There are entities or things («Personalities») that are the subject of an assertion. They are usually expressed by nouns.There are actions, processes, changes («Energy») that are not exclusively, but best expressed by verbs.There are properties and qualities that adhere to the entities or things and describe their state («Matter»). They are not exclusively, but best expressed by adjectives.The facets of Space and Time finally express circumstances that contextualise the assertion. They are not exclusively, but best expressed by the word type «adverb».

Let us illustrate this with an example. The assertion «Last week, Gill damaged her computer when she slipped on an icy sidewalk» can be broken down as follows (*cf.* Cumyn *et al.* 2018, at 884):

(1) Personality: Gill, computer
(2) Matter: icy
(3) Energy: slip, damage
(4) Space: sidewalk
(5) Time: last week

ad 1.: «Gill» and «computer» are nouns, and (in contrast to «sidewalk» and «week») they could be nothing but nouns due to their function, because their character as entities (in the broad sense of «personality») is crucial here. *ad* 2.: «Icy» is an adjective. *ad* 3.: To «slip» and to «damage» are verbs, and indeed the process or action («energy») of slipping and the associated damaging are crucial to the assertion. One could also write: «Gill's slipping damaged her computer». Then «slipping» would be a noun; however, this wording is somewhat artificial, because here a process is to be expressed. In addition, one also could say: «The computer is damaged.» Then «damaged» would be an adjective, but it would not be the same assertion, because the emphasis would no longer be on causation (action), but on the state of the computer. *ad* 4. and 5.: The assignment is self-explanatory. The words «sidewalk» and (last) «week» are nouns, but they are part of an adverbial phrase, whose function it is to describe the circumstances, not the core assertion.

Assuming the assertion «Last week, Gill damaged her computer when she slipped on an icy sidewalk» were to describe the factual basis for a claim in damages by Gill against the municipality, it would be indexed using the following concepts, according to our model (we refer in square brackets to terms drawn from our controlled vocabulary):

– *Person*: not applicable (there is nothing legally relevant to say about Gill)
– *Action*: fall [chute], breakage [bris]
– *Thing*: computer [ordinateur]
– *Context*: slippery surface [surface glissante], sidewalk [trottoir]
– *Legal category*: civil liability (public authority) [responsabilité civile (autorité publique)]
– *Sanction*: compensation damages [dommages intérêts compensatoires]

It is no coincidence if the six questions that journalists are trained to ask when covering a story (who, what, when, where, why, how) are reminiscent of the PMEST formula. In her best-selling coursebook on legal research, A. Sloan (2018, at 27) notes that users, when presented with a set of facts, tend to use the six journalistic questions to generate a list of search terms (*cf.* the popular formula in German legal education «Wer will was von wem woraus?»). It is also possible to categorize the relevant information by identifying the parties, places and things involved, the potential claims and defences, and the relief sought by the complaining party (Sloan 2018, at 27–28). Our model responds to the structure of legal research by ensuring that corresponding keywords are used to index legal documents, and that the search interface reflects a similar structure.

Comparing Ranganathan's scheme with the four functionally comparable facets of our model that describe the facts, it can be seen that we have refrained from giving the dimensions of time and space a prominent position and have instead created a single facet named «Context» for all circumstances that complete the action in accordance with the linguistic adverb function (if legally relevant). In contrast, we have divided the Personality facet of the PMEST formula into the two facets Person and Thing, because the distinction between persons (subjects) and things (objects) plays a decisive role in law. This view is confirmed by our observation that terms assigned to Person and Thing could not be anything other than nouns, unlike the terms of other facets which could easily be converted into verbs, adjectives and adverbs. Linguistically speaking we integrated the level of cases (casi) into the structure in addition to the word types: Person corresponds to the subject in the grammatical sense, Thing to the object. But here again, the comparison should not be taken too literally. In the sentence «A beats up X for no reason» X is grammatically seen an (accusative) object, nevertheless we would index the victim X not under Thing, but under Person. From a legal point of view, people are always subjects, not objects. However, we had considered distinguishing between perpetrators and victims and dividing the Person facet accordingly, but finally rejected this idea for practical reasons.

4. Potential for automatic indexing

A further step will be to consider whether our model may serve as a tool for the (semi-)automatic text-based and ontology-based keyword extraction and semantic indexing of cases. Both supervised and unsupervised data mining techniques can be considered.

There have already been numerous attempts to develop systems for the automatic indexing of (legal) texts (see *e.g.* Korycinski / Newell 1990; Francesconi *et al.* (eds.) 2010, Part 3; Gödert 2013; see also patent US 7,840,524 B2 of November 23, 2010). So far, this has not been achieved in a generally convincing way. Many established editors continue to use human indexing (albeit computer-assisted), despite its cost, because fully automated indexing does not yet meet their standards. Where the search process is humanly curated, it delivers better results (Nevelow Mart 2013, at 43, comparing Westlaw and Lexis).

However, we believe that our faceted indexing scheme offers an advantage over unstructured indexing in the process of automation. Faceted arrangement of knowledge elements can promote the modelling of semantic knowledge (Gödert 2014, at 129, 131). The existence of a semantic indexing scheme, organised by well-defined facets together with a controlled vocabulary, and the degree of normalization it ensures reduces ambiguities, reveals similarities among documents and could make indexing more predictable for the self-learning mechanisms of the system. The algorithm perhaps thus better knows what to look for in the document and which index term to use (or to recommend). The sample of 2,500 cases currently included in our prototype is probably not sufficient to support the machine learning required for automatic indexing. However, our detailed indexing policy would facilitate expansion of the existing sample by adding and manually indexing further cases.

In any case, the 2,500 sets of manually attributed index terms are valuable in themselves. They formalize the legal essence of the database content in a matrix (or template) that is not unlike a (light) legal ontology, revealing statistically analyzable information about the relationships between the terms and the concepts they represent, depending on the facets to which they belong (*cf.* Schweighofer 2009). Indeed, we intend to analyze this matrix using statistical tools in order to reveal patterns and correlations. For example, it is conceivable that the assignment of certain index terms to one facet will correlate significantly with the assignment of other terms to another facet. It is also conceivable that clusters of indexed terms will emerge, because they co-occur significantly. It is also planned

- to analyze the dataset according to a probabilistic topic modelling approach (for topic modelling of legal texts see *e.g.* Livermore *et al.* 2017),
- to compare its distinctive verbal pattern with our faceted scheme and with the analysis of the indexing matrix
- and possibly to create a supervised version of the used topic model combining it with the indexed dataset as a training corpus with multi-label data (*c.f.* Rubin *et al.* 2012, at 161–162).

Our first probabilistic topic modelling experiments were quite encouraging. Finally, it might be interesting to compare our model using humanly predefined (static) facets with the approach of Dakka and Ipeirotis, who experimented with the extraction of automatically generated facets from newspaper articles (Dakka / Ipeirotis 2008). Such findings could be incorporated into an algorithm for (semi-)automatic indexing.

Acknowledgments. This research program is funded by the Social Sciences and Humanities Research Council of Canada (2015–2020). We are grateful for the research assistance of Christophe Achdjian, Dominique Boulanger, Cécile Gaiffe, Marjorie Gauchier, Joanie Gosselin-Lepage, Ariane Joncas, Béatrice Lecomte, Amélie Roy, Erin Sandberg, Forrest Wakarchuk and Iryna Zazulya. Thanks to Alexandre Fortier, our expert librarian, David Lesieur, our programmer, and Charles Tremblay-Potvin, our project coordinator. Finally, we would like to thank the Société québécoise d'information juridique (SOQUIJ) for making their data available to us.

5. References

Alexy, Robert (2003). On Balancing and Subsumption. A Structural Comparison. Ratio Juris. Vol. 16: 433–49.

Birks, Peter (1997). Definition and division: a meditation on *institutes* 3.13. In: Birks, P. (ed.), The Classification of Obligations. Clarendon Press, Oxford: 1–35.

Broughton, Vanda (2001). Faceted classification as a basis for knowledge organization in a digital environment; the Bliss Bibliographic Classification as a model for vocabulary management and the creation of multi-dimensional knowledge structures. The New Review of Hypermedia and Multimedia 7(1): 67–102.

Broughton, Vanda (2010). The use and construction of thesauri for legal documentation. Legal Information Management 10(1): 35–42.

Broughton, Vanda (2011).). Brian Vickery and the Classification Research Group: the legacy of faceted classification. https://www.researchgate.net/publication/266184326, last visit January 31st, 2019.

Classification Research Group (CRG) (1955). The need for a faceted classification as the basis of all methods of information retrieval. UNESCO Ref. 320/5515 of May 26, 1955, https://unesdoc.unesco.org/ark:/48223/-pf0000179561, last visit January 31st, 2019.

Cumyn, Michelle (2015). The Structure of Stateless Law. In Shauna Van Praagh and Helge Dedek (eds)., Stateless Law: Evolving Boundaries of a Discipline, Ashgate, ch. 7.

Cumyn, Michelle / Gosselin, Frédéric (2016). Les catégories juridiques et la qualification : une approche cognitive. 62-2 Revue de droit de McGill 329-387, http://id.erudit.org/iderudit/1040050ar, last visit December 22th, 2018.

Cumyn, Michelle / Hudon, Michèle / Mas, Sabine / Reiner, Günter (2018). Towards a New Approach to Legal Indexing Using Facets. In: Mouhoub, Malek *et al*. Recent Trends and Future Technology in Applied Intelligence. IEA/AIE 2018 Montreal: 881–888.

Dakka, Wisam / Ipeirotis, Panagiotis G. (2008). Automatic Extraction of Useful Facet Hierarchies from Text Databases. https://www.researchgate.net/publication/4331052, last visit December 22th, 2018.

Farradane, Jason (1979). Relational Indexing. Part I. Journal of Information Science 1 (1980): 267-276.

Francesconi, Enrico / Montemagni, Simonetta / Peters, Wim / Tiscornia, Daniela (eds.) (2010). Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language. Heidelberg.

Gerathewohl, Peter (1987). Erschließung unbestimmter Rechtsbegriffe mit Hilfe des Computers: ein Versuch am Beispiel der «angemessenen Wartezeit» bei § 142 StGB, doctoral thesis, Universität Tübingen.

Gödert, Winfried (2013). An Ontology-based Model for Indexing and Retrieval. https://arxiv.org/ftp/arxiv/papers/1312/1312.4425.pdf, last visit December 22th, 2018.

Gödert, Winfried (2014). Facets and Typed Relations as Tools for Reasoning Processes in Information Retrieval. S. Closs et al. (Eds.): Metadata and Semantics Research 2014, CCIS 478: 128–140.

Godbold, Natalya (2006). Beyond information seeking: towards a general model of information behaviour. Information Research, 11(4) paper 269, http://InformationR.net/ir/11-4/paper269.html, last visit December 22th, 2018.

Gross, Tina / Taylor, Arlene G. / Joudrey, Daniel N. (2015). Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching. Cataloging & Classification Quarterly, 53:1–39.

Hearst, Marti A. (2006). Design Recommendations for Hierarchical Faceted Search Interfaces. In: Proc. SIGIR 2006, Workshop on Faceted Search: 26–30.

Hudon, Michèle / El Hadi, Widad M. (2017). Introduction. La classification à facettes revisitée. De la théorie à la pratique. Les Cahiers du numérique Vol. 13: 9–24.

ISO 25964-1-2011 (E). Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval. Geneva.

Komlodi, Anita (2002). The Role of Interaction Histories in Mental Model Building and Knowledge Sharing in the Legal Domain, Journal of Universal Computer Science, vol. 8, no. 5: 557–566.

Korycinski, C. / Newell, Alan F. (1990). Natural-language processing and automatic indexing, The Indexer Vol. 17: 21–29.

Krathwohl, David R. (2002), A Revision of Bloom's Taxonomy: An Overview, Theory into Practice, Vol. 41; 212 – 218.

Kuhlthau, Carol C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of the American Society for Information Science, 42(5): 361–371.

Kwasnik, Barbara H. (1999). The Role of Classification in Knowledge Representation and Discovery. LIBRARY TRENDS, Vol. 48: 22–47.

Ling, Justin (2018), Federal government looks to AI in addressing issues with immigration system, The Globe and mail, May 31, 2018, https://www.theglobeandmail.com last visit December 22th, 2018.

Livermore, Michael A. / Riddell, Allen B. / Rockmore, Daniel N. (2017). The Supreme Court and the Judicial Genre. 59 Ariz. L. Rev.: 837–901.

Maniez, Jacques (1999). Du bon usage des facettes. Documentaliste – Sciences de l'information 36(4/5), 249–262.

Moeller, Tony (1993). Optimierte juris-Nutzung mit den Mitteln der Rechtsinformatik unter Berücksichtigung überkommener juristischer Methodenlehre: zugleich ein Beitrag zur Abbildung des deduktiven Hauptschemas der analytischen Begründungslehre als Computermodell, doctoral thesis, Universität des Saarlandes, Saarbrücken.

Nevelow Mart, Susan (2013). The Case for Curation: The Relevance of Digest and Citator Results in Westlaw and Lexis. Legal Reference Services Quarterly, 32:1–2: 13–53.

Prestel, Bernhard M. (1971). Datenverarbeitung im Dienste juristischer Dokumentation: ein Arbeits- und Funktionsvergleich zweier Systeme. Berlin.

Ranganathan, Shiyali Ramamrita (1959). Elements of Library Classification. 2nd edition. Edited by B.I. Palmer. London: Asia Publishing House.

Reiner, Günter (2019). Juristisches Denken bei und für Juristen und Nichtjuristen: ein funktioneller Blick von außen, in: Kuhn T., Kramer U. and Putzke, H, (2019). Was muss Juristenausbildung heute leisten?, Stuttgart (forthcoming); 219–319.

Rouvroy, Antoinette (2018). Homo juridicus est-il soluble dans les données?, in: Droit, normes et libertés dans le cybermonde: liber amicorum Yves Poullet. Bruxelles: Larcier: 417–444.

Rubin, Timothy N. / Smyth, Padhraic (2012). Statistical topic models for multi-label document classification. Mach Learn 88:157–208.

Satija M. P. (2017). Colon Classification (CC). Knowl. Org. 44 (2017) No.4: 291–307.

Scott, Mark / Smith, Nigel (2010). Legal Taxonomy From Sweet & Maxwell. Legal Information Management 10(3): 217–222.

Schweighofer, Erich (2009). Learning and Verification of Legal Ontologies by Means of Conceptual Analysis. https://www.researchgate.net/publication/242641747, last visit December 22th, 2018.

Schweighofer, Erich / Lachmeyer, Friedrich, Trends und Communities der Rechtsinformatik, 2017, https://www.univie.ac.at/RI/IRIS2019/wp-content/uploads/2013/09/Schweighofer-ES-FL-Trends-Communities-fin.pdf, last visit December 22th, 2018.

Sloan, Amy E. (2018). Basic Legal Research: tools and strategies. 7th ed. New York.

Strauch, Hans-Joachim (2017). Methodenlehre des gerichtlichen Erkenntnisverfahrens: Prozesse richterlicher Kognition. München.