Artificial Intelligence and Law **7:** 341–361, 1999.

Advanced techniques for legal document processing and retrieval

E. PIETROSANTI[1] and B. GRAZIADIO[2]
[1]*FINSIEL S.p.A., Via Carciano, 4 - 00131, Roma, Italy*
E-mail: e.pietrosanti@finsiel.it
[2]*Via del Ponte di Piscina Cupa, 43 - 00128, Roma, Italy*
E-mail: b.graziadio@finsiel.it

Abstract. A large interest has been dedicated in recent years to the study of models for textual databases amenable to an effective integration of *search* and *navigation* functions. In the field of legal databases the need for sophisticated models is emphasised by the need to relate and combine in an effective way different types of texts, in order to solve legal problems.

In our research we have analysed several existing models, each providing specific benefits and exhibiting corresponding limitations, under both a functional and economical viewpoint.

Under a functional point of view, a distinctive feature of our model is the representation of relevant *context* information, aimed at improving the retrieval accuracy, in a framework in which the availability of multiple (structural, conceptual and functional) views over the legal texts emphasises the issues of the *transparency* of the model and of the *incrementality* of the search process. The model has been experimented on a significant excerpt of the Italian banking regulations and fiscal law, embodied in the NaviLex experimental system.

On the other hand, sophisticated models imply complex text encodings, which in turn entail high costs for the manual indexing/authoring task. This well-known problem, which hampers the development of large powerful systems, has been tackled with a set of specific linguistic tools, first experimented in the Esprit II project *Nomos* and subsequently developed in research and development projects carried out in the Finsiel Group. These tools – devoted to the automatic extraction from texts of the information structures considered in the retrieval model – use *shallow* techniques amenable to effective large-scale text processing in the legal domain, in order to overcome the state-of-the-art limitations of traditional 'deep' NLP techniques.

This article presents an overview of our approach, providing a general description of the representation model and processing tools, and concentrating primarily on the representation features and search improvements related to the use of the functional context information.

1. Introduction

The goal of most legal work – seen as a process of text handling – is actually to combine different types of texts in an effective way. Statutes, regulations, cases, precedents, legal literature, contracts are examples of documents that may have to be investigated together in order to solve a legal problem or even to be able to

understand the practical meaning of a legal rule. This well-known phenomenon can be described as *legal rule fragmentation*: the necessary information is often scattered in different documents or even in different data banks, and the links among the required pieces of information are difficult to establish. This problem, though particularly hard in a strongly text-centred field like the legal domain, is not limited to the legal area.

As a consequence, in the recent years a growing interest has been dedicated to the study of complex representation models, in which sophisticated *search* functions (typical of the Information Retrieval field) are integrated with *navigation* functions (typical of Hypertext systems). An effective combination of the benefits provided by the two models is expected to provide the best support tools for the localization of scattered information that is of interest for the user. The rapid expansion of the Internet, which has resulted in a rapidly growing worldwide hypertext, has provided additional momentum to the research in this area.

Among the different models that have been proposed in various fields, in the legal domain the interest for legal databases led to the adaptation for legal data (Agosti et al. 1991; Di Giorgi and Nannucci 1992) of a general two-level model (Agosti et al. 1991a) providing a conceptual layer intended to improve the system transparency. Although a sound conceptual layer is a vital component of any effective model, we believe that a functional limitation of many existing models is the lack of *context information*, suitable to be combined with concepts to improve the retrieval accuracy (especially the precision component).

Besides this functional limitation, an economical problem which hampers the development of powerful models is the cost for the manual indexing/authoring task, which is devoted to the extraction from texts of auxiliary data suitable to encode various relevant aspects of text content (typically, cross-reference citations and concepts belonging to a pre-defined thesaurus or classification schema). Besides being a hard and time-consuming task, this manual encoding activity is also error-prone and exposed to a substantial degree of subjectivity. In the legal field the importance of this problem can be easily understood by considering the extremely rapid growth of the overall document collection: more than one million new cases and statutes per year, according to Hafner (1990).

Natural Language Processing (NLP) techniques are a key resource in order to overcome this economical problem, but the state-of-the-art of NLP does not allow to envisage effective solutions to this problem on general domains. Nonetheless, in limited sub-languages (Kittredge and Lehrberger 1982) the feasibility of an automatic mapping from texts to suitable information structures has been demonstrated in various fields (Liddy et al. 1991; Rama and Srinivasan 1993). In the legal domain (Graziadio et al. 1992; Giannetti et al. 1992; Pietrosanti et al. 1995) have demonstrated the potential of ‘shallow’ NLP techniques, as opposed to the traditional ‘deep’ techniques, for effective large-scale text processing in the legal domain.

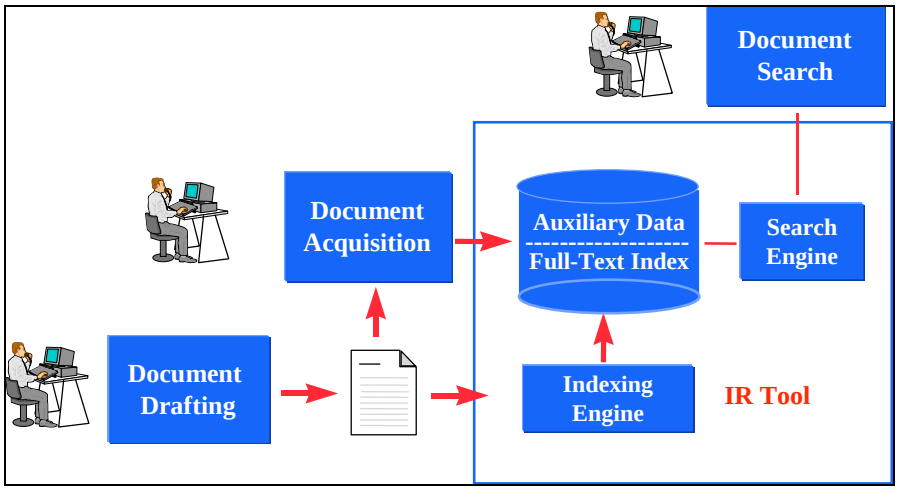

Figure 1. A framework for legal document lifecycle support.

In summary, the limitations of traditional approaches are related to two major drawbacks:

- the poor selectivity of the search criteria offered by traditional Information Retrieval systems does not allow the user to satisfy his/her information needs;
- the need to manually extract and encode the auxiliary information entails high costs and substantial risks of approximation and subjectivity.

Moving from the previous considerations, our work in the legal field has therefore conformed to the following guidelines:

- emphasis on the modelling of context characterization of words and concepts, to improve document search precision;
- use of semi-automatic tools for the acquisition of the legal information components;
- experimentation of the potential of the search and indexing techniques for the development of advanced document-drafting support.

In view of these guidelines, we have studied and developed a suite of specialized tools, intended to support the three fundamental steps of the lifecycle of a legal document (Figure 1):

- Acquisition (i.e., extraction and encoding) of auxiliary information from legal documents:
  – legal cross-reference citations (*RifLex*) – classification concepts (*ClassiLex*)
  – functional scheme views of documents content (*SchemaLex*)

- Intelligent search and navigation through legal databases (*NaviLex*)
- Legal document drafting support (*DraftLex*)

These tools – realized in various research and development projects, with experimentations in several legal domains related to different parts of Public Administration, including Public Account, Finance and Treasury – are described in the following sections of this paper, whose structure is as follows.

Section 2 describes the structural, conceptual and contextual dimensions of the reference retrieval model, together with the fundamental search and navigation functions made available by the model. The presentation provides also (2.4) a detailed overview and discussion of related approaches.

Section 3 is devoted to a description of the tools, starting from the search and navigation aspects (NaviLex) and then concentrating on the linguistic tools for the automatic extraction of the auxiliary information components of the model.

Section 4, devoted to some conclusive remarks, includes also an outline of the current research in the direction of legal drafting support.

2. Representation of the Structural, Conceptual and Functional Dimensions of Legal Documents

The proposed reference model for the content representation of legal documents is described in this section in terms of the multiple views (structural, conceptual and functional) provided over the legal texts.

The *structural dimension* of the content representation includes the hierarchical organization of legal documents and the web established by the legal cross-references, which can both be used for navigation purposes, allowing respectively direct access to a given text (through a table of contents) and a hypertext access through the ingoing or outgoing citations.

The *conceptual dimension* is based on the definition of complex linguistic terms (namely noun phrases) which constitute more effective content descriptors (as those typically included in the index of a book) with respect to word-based indexing (Croft et al. 1991; Evans et al. 1991).

Another crucial dimension of the content representation is the *functional component*, which allows associating to the concepts specific *functional roles*, which constitute meaningful contexts for the user (e.g., the definiendum of a definition, or the subject of an obligation).

2.1. THE STRUCTURAL DIMENSION

Legal databases exhibit a general structure characterized by the well-known subdivisions of legal documents (e.g., sections, articles and paragraphs in statutes; chapters, sections and sub-sections in regulations). In addition to constituting the basic text cohesion device, this hierarchical organization is an essential aspect of

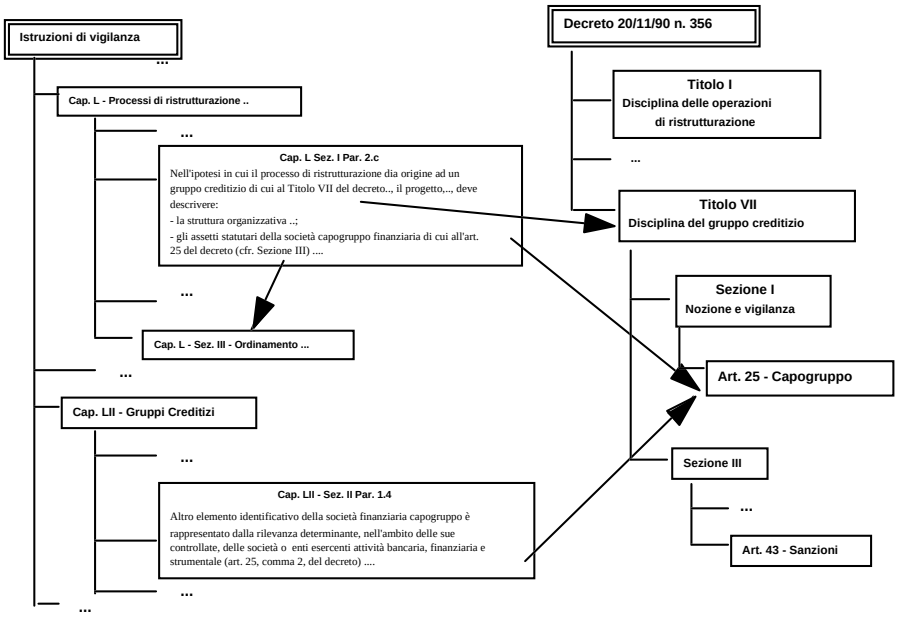

Figure 2. Structural dimension of the document base (regulation and statute).

the structural dimension of legal documents, as it provides also an 'address space' that is extensively used to make explicit reference, in a given legal document, to related parts of the same or other documents. The links established by these references make a complex cross-referring web of documents, that constitute the second component of the structural dimension.

As a pictorial description of this information, Figure 2 presents the components of the structural dimension in the NaviLex database, concerning the legal sub-domain of the Banking Legislation, which includes the *Decree* ("Decreto") represented in the right-side structure. The articulation is typical of Statutes (issued by the Parliament): the Decree is composed of numbered *Articles* ("Articoli") organized in *Titles* ("Titoli") that are in turn made of *Sections* ("Sezioni"), both indexed by roman numbers.

In order to specify and give the correct interpretation of the whole Banking Legislation – that is scattered in several Laws and Decrees – the national Bank of Italy (Banca d'Italia) issues the *Regulations on Bank Surveillance* ("Istruzioni di Vigilanza"). These Regulations are represented by the structure on the left hand side, that is articulated in Chapters ("Capitoli"), Sections and nested Paragraphs (for instance the upper text excerpt, that is discussed in detail in the following section, belongs to the Paragraph 2.c of the first Section of Chapter L). Due to the status of the Regulations, in addition to the internal cross-references pointing to other sections of the Regulations book, many external cross-references point the Decree, which contains the original legal rules referred to in the regulations. This

is a typical example of the need to navigate the cross-reference links to cope with the mentioned problem of the legal rule fragmentation.

2.2. PROVIDING A FUNCTIONAL CONTEXT FOR THE CONCEPTUAL INFORMATION: MODELLING NORM-KERNEL AND DEFINITIONS AS *functional schemes*

The investigation of legal concepts and their relations with other concepts is a crucial goal of much effort in legal work, aimed at identifying legally relevant items of knowledge and relevant relations between such items. The upper conceptual layer of the model (described in Figure 3) represents the universe of possible usable terms and their relationships. For the purpose of our model, a *concept* represents a meaningful entity for the domain. Each concept is linked to the documents, in which concept instances are denoted by *concept anchors*, that represent the linguistic manifestations of the concepts. In general, a set of possible linguistic expressions (i.e., the Concept Anchors in the documents) denotes a concept identified by a normalized linguistic term that represents the concept name (i.e., a noun phrase). For example, in the figure C_a-A_1 (e.g., "sales of goods") and C_a-A_2 (e.g., "to sell goods") are two anchors that identify two instances of the same concept C_a (e.g., "sale of goods")

Relations are established among concepts that are semantically linked. In the figure a generic hierarchical structure is depicted, that can be imagined as either a classification scheme or a complex thesaurus. Although we believe that the conceptual network is an important component for the complete model, in this paper the main focus is on the combination of concepts with relevant contextual information, aimed at improving the retrieval accuracy (especially the precision component).

The rest of the section is therefore devoted to the discussion of the contextual dimension, that is represented in the figure by particular frame structures. Since these structures express the particular function of the concepts in the context of the message communicated by the text, they have been named *functional schemes*.

The use of norm frames as a plausible method for the conceptual representation of legal knowledge has received large consensus in the legal theorist community. In van Kralingen et al. (1993) this model is discussed in the context of Legal Knowledge Based system, in view of the particular task of the support to drafting of legislation. Although their goal – i.e., the complete representation of a norm – is far beyond the scope of our work, the representation schema that we propose shares with their model the notion of *norm-kernel* (a long standing concept, cf. Von Wright (1963)). The *norm-kernel* is supposed to contain the essential information conveyed by a norm, answering questions like: who ought to do something?, what should he do?, etc. This leads to the consideration of the *legal modality*, *subject*, *object* and *conditions of applications* of a norm.

The *legal modality* determines the function of a norm, that is either an obligation (a command or a prohibition: *ought* or *ought not* respectively) or a permission

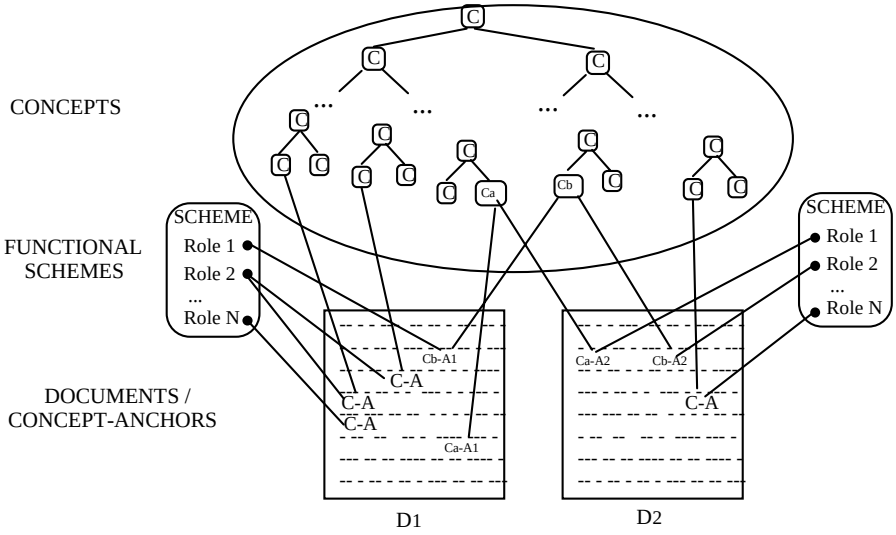

Figure 3. Conceptual and functional dimensions of the model.

(*may*). The *subject* of a norm is the person or institution to whom the norm is addressed. The *conditions of application* establish the circumstances under which a norm is applicable.

All of these components of the norm-kernel have been considered in the functional schemes of our model, because they play an important role as context qualifiers for the legal concepts. In the mentioned complete models for legal knowledge representation, the *object* of a norm is a fine-grained description of the act which falls under the scope of the norm, taking into account details such as the modality of the action, the setting of the action (e.g., spatial, temporal and circumstantial aspects) and the rationale of the action (causality, finality, intentionality). We take a different approach, and reduce the complex activity components to a couple of slots: the *action* – i.e., the activity performed by the subject of a norm – and the *object* , that can be either a direct or indirect object of the action. In addition, we define a generic *context* role to take into account both the conditions of applications and the possible aspects of the action (e.g., temporal and spatial aspects).

We have considered so far the essential component of the content of legal texts, namely the propositions (norms) that express general rules, standard of behaviour and principles.

Another fundamental component of legal texts is represented by the *definitions* of legal concepts. The knowledge of the definition(s) of legal concepts is an essential pre-requisite for the correct interpretation of norms. In our model the definitions are represented by simple schemes that are intended to capture the essential functional roles played by the concepts that appear in the definitions. Using traditional naming conventions, we have thus identified the following functional roles: *definiendum* (the role played by the concepts that are defined in the

definition); *definiens* (the role of the concepts that appear in the definition body); *definition context* (analogous to the corresponding *context* role for norms, generally includes concepts related to conditions of applications).

Following the previous considerations, our model takes into account the following fundamental functional schemes:

SCHEME TYPE: OBLIGATION/PERMISSION

Subject
⟨subject of the obligation/permission⟩
Action
⟨activity performed by the subject of a norm⟩
Object
⟨object of the action⟩
Context
⟨relevant concepts for the context of applicability⟩

SCHEME TYPE: DEFINITION

Definiendum
⟨concept(s) to be defined⟩
Definiens_
⟨defining concepts⟩
Definition context
⟨relevant concepts for the conditions of applicability⟩

The choice of simple one-level functional schemes has a twofold motivation. First, our goal is to provide a simple conceptual and functional view of the document content, to be directly used for intelligent retrieval purposes. Complex schemes with a large number of nested slots would be difficult to manage for the user. The second motivation is related to the strategic goal of automatic acquisition of the content representation. While we have experimented (as shown in Section 3.2) the feasibility of shallow NLP techniques capable of extracting the simplified schemes directly from texts, the automatic acquisition of a fine-grained representation is hampered by the state-of-the art limitations of full-fledged NLP.

Generally speaking, our model can be situated at the knowledge representation level which has been named in Nanard et al. (1993) *macroscopic semantics*, being far more detailed than simple indexing by weighted keywords, but far less detailed than a conceptual-graph based description needed for complete text understanding.

To clarify the previous considerations, we present (left) a brief excerpt extracted from Chapter L, Section I, Paragraph 2.c of the mentioned Regulation, and (right) the corresponding representation in terms of concepts assigned to the relevant roles of a functional scheme. The relevant concepts are highlighted in bold, while the underscored words emphasize the recurrent cue phrases that are used by the acquisition modules of SchemaLex (see Section 3.2 below).

"*Should* the ***reorganization process*** give origin to a ***group of credit institutions*** as mentioned in ***Title VII of the decree*** . . . , the ***project***, . . . , *ought to* ***describe***:
– the ***organizational structure*** . . . ;
– the ***articles of association*** of the ***company*** as mentioned in art. 25 of the decree (see Section III) . . . "

SCHEME TYPE: OBLIGATION
Subject
project
Action
describe
Object
organizational structure, article of association, head holding company
Context
reorganization process, group of credit institutions Title VII of the decree

2.3. BASIC SEARCH FUNCTIONALITIES USING FUNCTIONAL SCHEMES

In order to clarify the potential of the representation model, we outline the fundamental functions provided by the model for search and navigation support that have been included in NaviLex, concentrating on the functions related to the conceptual and contextual dimensions.

The functions related to the conceptual and contextual dimensions make reference to the following basic entities of the model previously outlined (see Figure 3):

- concepts
- concept anchors (instances)
- functional schemes and roles

- **Context-based concept browsing**
  Given: 1) a document set DS (possibly the entire document collection);
  2) a functional role FR

The function returns a set of concepts whose instances play a role FR in some document belonging to the document set DS.

For example, given the two-document collection of Figure 3 and the functional role Role 1, the function returns the concept set $\{C_a, C_b\}$. In case of a definition scheme (Role 1 = Definiendum) $\{C_a, C_b\}$ would be the set of concepts for which a definition is present in the given document set.

- **Document selection based on the functional role of a concept**
  Given: 1) a document set DS (possibly the entire document collection);
  2) a Concept C;
  3) a functional role FR

The function returns the document subset of DS that include instances of the concept C playing the functional role FR. For example, given the two-document

collection of Figure 2, the concept C_a and the functional role Role 1, the function returns the document set {D2}.

2.4. RELATED WORK ON HYPERTEXT RETRIEVAL MODELS

The purpose of this section is to discuss the distinctive features of our approach in comparison to other models and systems, that have in part been mentioned in previous sections. As observed in Arents and Bogaerts (1993), all the models recently presented have in common the separation between the *document space* (the documents in the hypermedia system) and the *index space* (the indices that characterize these documents).

Our model conforms to this general viewpoint and in particular can be closely related to the *EXPLICIT* model presented in Agosti et al. (1991) and referred to in Di Giorgi and Nannucci (1992), where the document and index space are respectively indicated as *hyperdocument* and *hyperconcept*. The hyperdocument, defined as a network of structural links combined with the network of reference links, corresponds closely to the structural dimension of our model. The conceptual dimension of our model matches the purpose of the hyperconcept, whose task is to handle the semantic structure of concepts used to describe the contents of document collection. The use of a rich semantic structure (including *indexing links* connecting thesaurus nodes to the documents as well as *classification links* used to aggregate documents according to classification criteria) is proposed also in Aboud et al. (1993).

With respect to these models, a crucial additional feature of our model is the consideration of the context information, provided by the functional schemes (contextual dimension).

The contextual information is taken into account in Arents and Bogaerts (1993) in connection to what they call *semantic hyperindices*, that rely heavily in the use of thesauri to support browsing search. The example "*Definition of* pitting.corrosion *of* titanium" shows how they use the context "definition" to qualify the occurrence of the concept "pitting.corrosion of titanium". In their work much emphasis is on a sophisticated description of concepts, which are embedded in a network that can be traversed both vertically and horizontally with a link navigation mechanism called 'broad-button'.

In Nanard et al. (1993) a model is described that explicitly considers the qualification of concepts by means of contexts (examples of contexts are "definition", "general rule", etc.), thus allowing contextual access to technical documents.

With respect to the mentioned models that take into account the context information, a distinctive feature of our model is that the functional schemes are actually *structured contexts*, in which the user can specify not only a context type (for example the "definition" scheme type) but also the functional roles associated to the concepts of interest.

Our notion of structured contexts is also related to the idea of *segmented database* that is discussed in Rau and Jacobs (1991), where the differentiation of keywords into segments allows to distinguish – in a constrained domain of commercial news – companies mentioned in passing from those actively involved in mergers or other events, and locations of companies from location of stories. This capability, that is achieved using sophisticated Natural Language Technology, is analogous to the possibility to distinguish (see the following Section 3.2) between the occurrence of a concept as the *definiendum* of a definition and the occurrence as the *subject* of an obligation.

In view of our research goal, related to the extraction and use of context information, important insights and results are given by Rama and Srinivasan (1993) who have reported on an investigation carried out on medical abstracts, in order to show how the qualification of keywords with their conceptual roles in a text can be used to derive a meaningful text-representation scheme. In addition to the text-grammar approach based on observable regularities in the structure of documents, that is related to our techniques for the recognition of functional schemes (see Section 3.2 below), their paper presents interesting experimental results. They have studied the role distribution of keywords, and found out that keywords exhibit role variation across abstracts, making the claim that this variation can be potentially exploited to make retrieval more precise. This result supports our experience with the use of functional schemes in the NaviLex system.

3. Advanced Tools for Legal Text Processing and Retrieval

In order to validate the adequacy of the model, we first designed and developed a prototype (*NaviLex*: *Navi*gation on *Lex*) addressed to a specific user, in the context of the banking legislation. The target users were expert legal drafters working in the legislative department of the Bank of Italy (Banca d'Italia), that are in charge of the drafting and maintenance (with respect to the evolution of the relevant statutes) of the regulations.

The first prototype (initially developed using Toolbook™, Pietrosanti et al. (1994) has then evolved into a system (developed in Windows/Visual Basic environment and based on the *Fulcrum Search Tools*™ Information Retrieval engine) which includes also a fiscal database (Value Added Taxation, V.A.T.).

In Section 3.1 we describe NaviLex, while Section 3.2 outlines the fundamental features of the linguistic tools aimed at the acquisition of the auxiliary data of the model.

3.1. SEARCH AND NAVIGATION OVER LEGAL TEXTUAL DATABASES WITH NAVILEX

The capability to use the functional schemes in order to enhance the search precision constitutes a distinctive feature of NaviLex. For instance, the user could retrieve the document containing the following definition:

> The *holding company* of a credit group is the company fielded in Italy which holds control of at least a *credit company* or a finance company ...

by searching for all the definition schemes in which the concept "holding company" plays the role of *definiendum (*ruling out all the other documents in which "holding company" occurs only incidentally) or, alternatively, searching for all the definitions which use the concept (role *definiens*) "credit company".

The representation model adopted in NaviLex is general and flexible, as it can be used at different levels of representation of legal information. The previous example illustrates the combined use of functional roles and keywords, while the next example describes a use of functional schemes integrated with the usual full-text search typical of IR systems (which does not require a previous document classification) . The snapshot in Figure 4 illustrates the result of a common full-text search using a word-truncated pattern (intended to represent some variants of the term 'prestazione di servizi' – *services*). The result list is composed by 19 documents, sorted in order of statistical relevance, computed according to a vector-space model similarity function (Salton and McGill 1984).

In addition to the 'linear' hit-list, the system can visualize also a Hierarchical View (*Vista Gerarchica*) of the result list, based on the structural dimension of the database, which depicts a hierarchical aggregation of the documents, related to their collocations in the specific legal sub-domain. The numbers associated to the hierarchy nodes (folders) give the number of retrieved documents underlying the node: for instance, the figure contains reference to two documents of the Bank of Italy sub-domain (BKI) and 17 of the V.A.T. (IVA) domain, which in turn are split among legislation (7) and cases (10).

The two views are synchronized, and allow the user to select, for instance, a document in the hierarchical view (e.g., art. 4 of DPR 633), and see the collocation of the same document in the sorted result list.

Let's now suppose that for the user the result list contains too many documents, so he decides to make the search criteria more selective, for instance by concentrating on the documents in which the required terms occur only as part of a definition.

The new query can be formalized by selecting the value 'definizione' (*definition*) for the field '**Tipo di Schema**' (*Scheme Type*). The corresponding result-list (see Figure 5) contains now only 8 documents. An interesting consequence of the new query formulation is that now two documents belonging to the BKI domain, previously ranked low due to statistical criteria, have been drawn to the user's attention due to their appearance inside some definition.

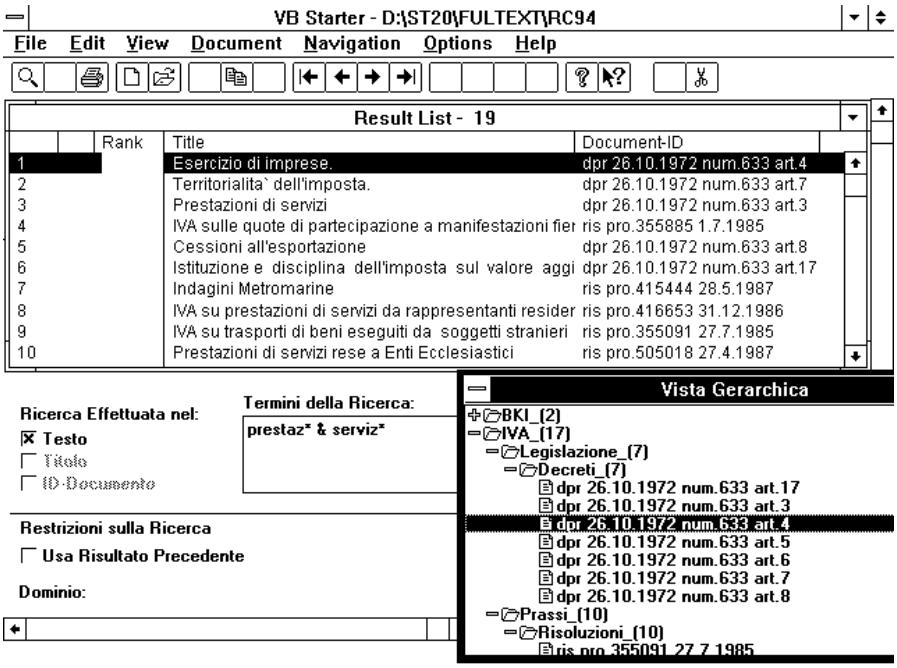

Figure 4. An example of simple full-text search with NaviLex.

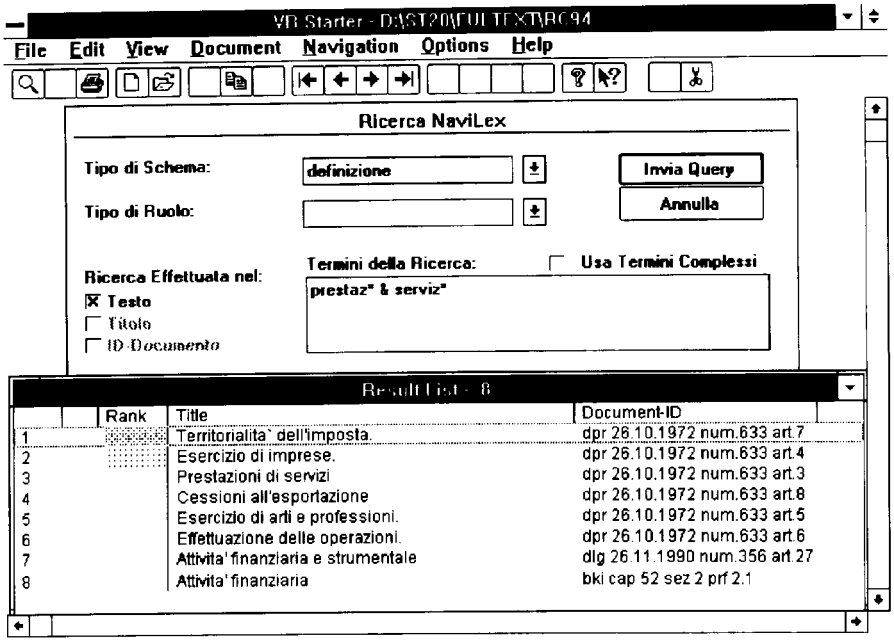

Figure 5. Full-text search with functional scheme restriction.

A further possible refinement of the query could lead to the retrieval of the documents in which the desired terms occur in some definition as (‘**Tipo di Ruolo**’ ‘*definiendum*’) the object of the definition.

In that case, only two documents would be retrieved, containing the two definitions of ‘prestazione di servizio’ existing in the database.

An interesting feature of this retrieval model is related to performance issues, due to the use of a native indexing feature available in the Information Retrieval engine *Fulcrum Search Server*^{TM}, namely the capability to define “Text Zones” that can be used to set additional “contextual” constraint to the search engine. For instance, in the NaviLex index schema specific zone types (e.g., the “def_definiendum” zone) have been defined, and in the documents to be indexed instances of these zones (e.g., “The *holding company* of a credit group” in the previous example) have been delimited by means of appropriate markup symbols, that are interpreted by the indexing engine.

Whenever the user wants the search terms to play the role of ‘*definiendum*’, the search algorithm will build a search statement (in SQL-like syntax) which contains the following clause:

. . . WHERE *def_definiendum* **CONTAINS** . . . ⟨*search terms combination*⟩

3.2. LINGUISTIC TOOLS FOR THE AUTOMATIC ACQUISITION OF THE AUXILIARY INFORMATION COMPONENTS

The current state of the art in NLP prevents these techniques from being used on full-text and/or unrestricted domains, due to computational complexity and ambiguity resolution problems.

If considered in the domain of legal texts, these problems exhibit peculiar aspects that provide additional challenges for the Natural Language Understanding task: legal documents are notoriously extremely complex, as they include long and ambiguous sentences with many cross-links as well as anaphoric references.

Despite the mentioned problems, legal texts present interesting opportunities as long as they express a tacitly defined ‘structural sublanguage’, exhibiting that special kind of regular structures that implicitly defines the ‘juridical style’. Legal texts are written according to fixed conventions, some of them explicitly stated and recommended, others just *de facto* standards: typographical layout, formal and recurrent expressions and a specialized vocabulary make legal texts an ideal testbed for experiments in the “sub-language” area (Kittredge and Lehrberger 1982; Liddy et al. 1981; Rama and Srinivasan 1993).

The issue of automatic knowledge acquisition from legal texts, relying on their highly structured nature, has been a major goal of the aforementioned Esprit II Project n.5330 “NOMOS: Knowledge Acquisition for Normative Reasoning Systems”.

The tools and techniques first experimented in the NOMOS project have been subsequently refined and extended in research and development projects (partially funded by the Finsiel Group joint research programme).

A specific set of linguistic tools has been devised in order to extract a 'shallow' content representation layer, including the structural, conceptual and functional components of legal texts outlined in previous sections. A crucial point is that in this case the relevant automatic acquisition modules rely on 'partial' NLP techniques, that do not require the most powerful, computational aspects of NLP, but are based on recurrent legal text peculiarities that make them suitable to be applied to large text databases.

In this framework the ***SchemaLex*** tool, aimed at the extraction of recurrent functional schemes, relies upon the encoding of recurrent textual patterns. For instance, the following:

1. "Per" X "si intende" Y (e.g.: "Per *domicilio* si intende *il luogo in cui si trova la sede legale*")
2. "Si considera" X Y (e.g.: "Si considera *domicilio il luogo in cui si trova la sede legale*")

correspond to typical definition patterns, and permit the identification of the functional role of the text segments X and Y, respectively playing the role of *definiens* and *definiendum*, resulting into the following target representation structure:

Segment	**Role**	**Value**
X	*definiendum*	"Domicilio" (*domicile*)
Y	*definiens*	"Il luogo in cui si trova la sede legale" – *the location of the registered office*

In addition to the basic segmentation criteria based on the existence of recurrent separation expressions, some heuristics can be used in order to separate contiguous segments. For instance, in the second pattern for definitions, the splitting point can be identified by interpreting the whole string as a NP + NP (noun phrase) sequence, recognized either by searching for some closed-class splitting words such as determiners and pronouns or by analysing the sentence with a shallow syntactic analyser. The type of analysis performed by SchemaLex is *partial* and *shallow*, as it relies on the knowledge of a relatively small number of textual elements. For instance, in order to analyse the aforementioned textual schemes, the system in general should not analyse the content of the X and Y segments (except for the possible search for clues indicating the boundary between X and Y) but simply tries to match patterns based on the presence of "per" (*for*) and of particular modes and tenses of definitory verbs such as "si intende" (*is understood*).

As already mentioned, in legal textual databases documents are usually classified by means of a set of keywords belonging to a pre-defined thesaurus or

classification scheme. The human indexer who is in charge of the task uses both a purely linguistic competence – allowing for the recognition of different morpho-syntactic variants of thesaurus terms – and a semantic competence related to a deep domain knowledge.

For instance, in relation to the thesaurus entry "cessione di beni" (*transfer of goods*) the following variants are likely to occur in a document:

- morphological e.g., "cessione di beni" vs. "X cede un bene" (*X transfers a good*)
- syntactical e.g., "cedere beni" vs. "beni ceduti" (active-passive transformation)
- semantical e.g., "cessione di beni" vs. "compravendita di beni" (synonyms)

In this framework, the ***ClassiLex*** tool is aimed at supporting the classification task, by finding out in the documents significant terms – which are related to some thesaurus term – that can be suggested to the user as candidate classification term. The linguistic variants managed in ClassiLex are essentially of the first two types (including nominalization phenomena: e.g., the expression in a text "... the treasury bonds bought by ..." can be automatically translated into the normalized form "purchase of treasury bond"). In order to give an effective support to the user, the thesaurus terms that the system suggests to the user should be accurately selected. To this end, the linguistic analysis modules used in ClassiLex can assign to each candidate term a couple of values indicating the *reliability* and *relevance* of the term. These indexes can be used in order to select and sort a subset of the extracted terms to be proposed to the user.

As illustrated in previous sections, legal cross-references constitute an essential component of the structural dimension of legal databases, that is usually extracted and encoded manually in order to perform search and navigation functions. The objective of the ***RifLex*** module is to analyse legal texts in order to find and encode the normative references. A Graphical User Interface (in Windows environment) allows the visualization and validation of the results, and the resolution of the possibly ambiguous solutions provided by the system. The recognition of the cross references spans the maximum granularity level of the legal documents' structure (namely, sub-partitions of paragraphs identified by letter/number), using techniques of 'partial' analysis, meaning that the scope of the processor is limited to the linguistic expressions which denote references. Anyway, these techniques are not 'shallow', as the artificial sub-language of normative references is affected by typical natural language phenomena which require sophisticated analysis techniques, aimed at solving complex problems such as incompleteness (ellipsis), anaphora (e.g., "fourth paragraph of the aforementioned article"), relative references (e.g., "preceding letter", "last article") and discontinuous constituents. An extensive description of the acquisition tools is beyond the scope of this paper. In order to highlight the main characteristics of the approach to the problem,

we discuss some details concerning RifLex (additional details can be found in Pietrosanti et al. (1995b).

The complexity of the sub-language is managed in RifLex by adopting a multi-layer processing strategy, aimed at analysing linguistic objects of increasing complexity, using a chart-parser designed and developed in Finsiel.

A particularly relevant problem is raised by the frequency of *incomplete* references, in which some necessary constituents are missing, and should thus be derived from the context. In these cases generally (but not necessarily) anaphoric references are present, and the missing information (the 'complement' of the incomplete reference) can be derived from preceding references, as in the following example:

> . . . as in the aforementioned article 20, paragraphs 2 and 3 . . .

whose *complement* (i.e., the indication of the regulation that includes article 20) can be extracted from a previous complete reference:

> . . . article 20, first paragraph, of the regulation passed with decree 25 may 1895, n. 350

The example lends itself to a fairly accurate analysis, thanks to the existence of the anaphoric reference "aforementioned article 20". In other cases, the correct solution requires some knowledge of the general context, as in the following example, in which the complement is constituted by the norm itself which includes the text:

> (Article 128) . . . for the completion of the services described by the previous article 127

In the general case, a list of possible complements should be taken into consideration, identified on the basis of various heuristics not mutually exclusive. This situation leads to an ambiguous set of likely solutions, which require the user's validation.

The current implementation runs in MS-Windows™ environment, and is made of two components: the first, developed using LPA-Prolog™, is the linguistic analysis module; the second one, developed in Visual Basic environment, implements the validation interface.

The upper window (Figure 6) contains the text to be analysed, the lower window presents the encoding of the current reference (that is highlighted in the text window).

The current reference in the figure is a *complex* reference, as it includes a list of two complex partitions (which are in fact distributed over two rows of the visualization grid), and is also *incomplete*, as a reference to the enclosing norm is missing. In this case the resolution of the ellipsis is done using the general context, and the record fields proposed for the completion (that are highlighted with a different colour) are proposed for the user validation. The user is informed about the number of the possible sources for the ellipsis resolution, and can also

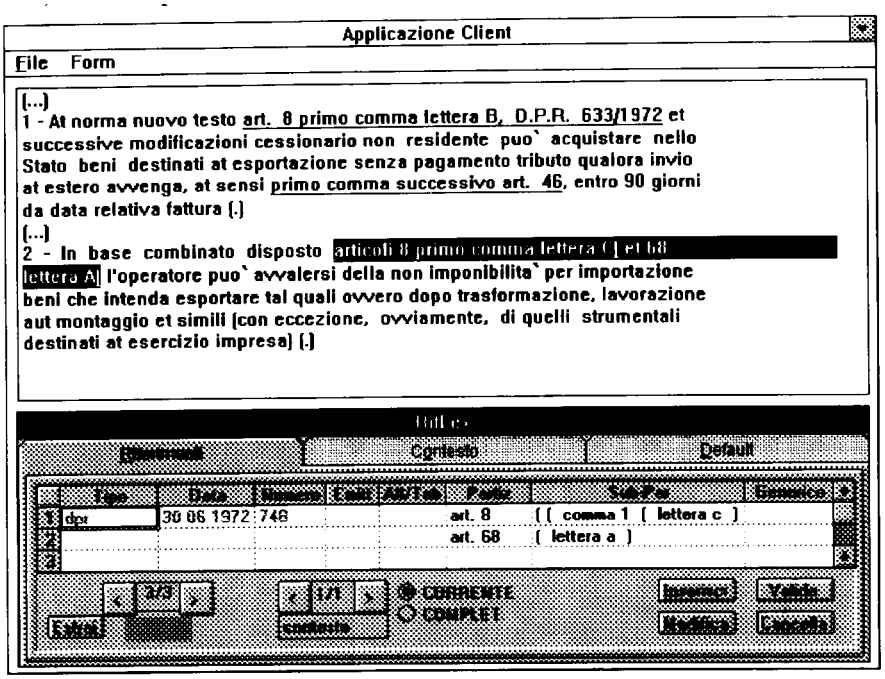

Figure 6. The RifLex user interface.

point to the source reference in the upper window. When necessary, the user can also modify the encoding proposed by the system, or delete the reference. Finally, by selecting the tabs “Contesto” e “Default”, the user can modify the values of the “Context” and “Default” information.

4. Conclusions and Future Developments

Starting from a critical evaluation of the benefits and limitations (under both a functional and economical viewpoint) of the existing models which integrate *search* with *navigation* functions, we have presented a reference model which takes into account *context* information in order to improve the retrieval accuracy, and has been tested in the NaviLex experimental system, dealing with a significant excerpt of the Italian Banking Regulations and V.A.T. fiscal law. Further evaluation is planned, aimed at evaluating the benefits due to the functional schemes retrieval, according to the methodology sketched in Rama and Srinivasan (1993).

We have also presented a suite of tools – devoted to the automatic extraction from legal documents of the information structures needed for advanced information retrieval purposes – using ‘shallow’ techniques amenable to effective large-scale text processing.

Future developments are planned in the following directions, matching the guidelines stated in Section 1.

Representation model. We are planning an accurate evaluation of the possible adoption of XML for the representation of the information (especially the structural and functional dimensions) of legal documents. The XML family of standards (XML 1998) is being defined by the WWW Consortium to ease the task of exchanging, manipulating and reusing document data, by maintaining a clear separation between the ways in which data, structure (DTD) and layout (stylesheet) are encoded. The availability of this standard representation framework should also provide suitable standard techniques for querying structured information, thus enhancing the search capabilities of NaviLex.

Linguistic Processing Tools. Functional extensions of the acquisition tools are aimed at obtaining a wider coverage of the linguistic phenomena and of the terminology on several domains. Major efforts are also aimed at re-engineering the tools, in order to make them suitable to be integrated in standard system architectures. In this framework, we are currently designing an evolution of the RifLex tool to be developed in C-Language.

Document Drafting Support. An important development direction is aimed at the extension and adaptation of the text-processing and search tools in a framework of Legal Drafting support. The objective is to integrate within commercial text-editors specific functionalities, which can help the legal drafter in two crucial tasks: searching all the existing legal documents which have potential relations with the new text he is currently writing; enforcing linguistic clarity, uniformity and coherence for the new text, by testing the application of drafting rules concerning various formal and substantial aspects of legal texts.

A first prototype version of the *DraftLex* system has been realized, in which specific functions are added to the MS Word™ editing environment.

In DraftLex the first type of support (sometimes named *cognitive* support) is provided by an extension of NaviLex (for instance, the user can select a concept in the document he is writing, obtaining all the existing definitions or obligations concerning this concept).

The second class of functions (*linguistic* support) is provided by integrating in DraftLex the linguistic analysis tools, used for correctness-checking purposes. For instance, an extension of RifLex is used to check the cross-references correctness, and the kernel of ClassiLex allows checking the terminology uniformity against a pre-defined thesaurus.

References

Aboud, M., Chrisment, C., Razouk, R., Sedes, F., and Soule-Dupuy. C. 1993. Querying a hypertext information retrieval system by the use of classification, *Information Processing & Management* **29**(3), 387–396.

Agosti, M., Colotti, R., and Gradenigo G. 1991. A two-level hypertext retrieval model for legal data, *Proceedings of the 14th International Conference on Research and Development in Information Retrieval*, SIGIR'91, ACM.

Agosti, M., Gradenigo, G., and Marchetti, P.G. 1991a. Architecture and functions for a conceptual interface to very large online bibliografic collections, in *RIAO 91, Intelligent Text and Image Handling*, Barcelona, April 1991.

Arents, H.C. and Bogaerts, W.F.L. 1993. Concept-based retrieval of hypermedia information: from term indexing to semantic hyperindexing, *Information Processing & Management* **29**(3), 373–386.

Croft, W.B., Turtle, H.R., and Lewis, D.D. 1991. The use of phrases and structured queries in information retrieval, *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Di Giorgi, R.M. and Nannucci, R. 1992. Hypertext systems for the law. In *Proceedings of International Conference Informatique et droit/Computers and Law*, Montreal 30 Sept–3 Oct.

Evans, D.A., Ginther-Webster, K., Hart, M., Lefferts, R.G., and Monarch, I.A. 1991. Automatic indexing using selective NLP and first-order Thesauri. In *RIAO 91, Intelligent Text and Image Handling*, Barcelona, April.

Giannetti, A., Dassovich, P., Marchignoli, G., Mussetto, P., Pietrosanti, E., Azzam, S., Celnik, P., Bilon, J., Fortier, V., and Pires, F. 1992. NOMOS: knowledge acquisition for normative reasoning systems. In L. Steels and B. Lepape (eds.), *Enhancing the Knowledge Engineering Process: Contributions from Esprit*, Elsevier Science Publishers.

Graziadio, B., Mussetto, P., Pesce, E., Pietrosanti, E. 1992. A multi-layered architecture for automatic knowledge acquisition from legal texts. In *Proceedings of 12th International Conference on A.I., E.S. and N.L.*, Avignon 92, 1–6 June.

Hafner, C.D. 1990. Challenges for text-based intelligent systems, *Proc. of AAAI Spring Symposium Series: Text-Based Intelligent Systems*, March 27–29, Stanford University.

Kittredge, R. and Lehrberger, J. (eds.) 1982. *Sublanguage: Studies of Language in Restricted Domain*, De Gruyter, Berlin.

van Kralingen, R., Oskamp, E., and Reurings, E. 1993. *Norm frames in the representation of laws*, in Svensson, Wassink and van Buggenhout (eds.), *Legal Knowledge Based Systems: JURIX '93: Intelligent Tools for Drafting Legislation*, Computer-supported Comparison of Law

Liddy, E.D., Jourghenson, C.L., Sibert, E., and Yu, S. 1991. Sublanguage grammar in natural language processing for an expert system. In *RIAO 91, Intelligent Text and Image Handling*, Barcelona, April.

Nanard, J., Nanard, M., Massotte, A., Djemaa, A., Joubert, A., Betaille, H., and Chauchè, J. 1993. Integrating knowledge-based hypertext for task-oriented access to documents, *Proceedings of the 4th International Conference on Database and Expert Systems Applications*, DEXA '93, Springer-Verlag.

Pietrosanti, E., Mussetto, P., Marchignoli, G., Fabrizi, S., and Russo, D. 1994. Search and navigation on legal documents based on automatic acquisition of content representation, *Proceedings of the Conference: RIAO 94: Intelligent Multimedia Information Retrieval Systems and Management*, New York, 11–13 October.

Pietrosanti, E., Mussetto, P., and Marchignoli, G. 1994b. *NaviLex: Integrating Search and Navigation in a Legal Hypertext based on Semi-Automatic Content Acquisition*, Informatica e diritto, No. 2/94 (IDG-Firenze) – special issue on *Hypertext and Hypermedia in the Law*.

Pietrosanti, E., Dassovich, P., Giannetti, A., Marchignoli, G., and Mussetto, P. 1995. *Automatic Knowledge Acquisition from Legal Texts: An Isomorphic Approach*, Proceedings of the Conference Towards a Global Expert System in Law, CEDAM 95 (Italy)

Pietrosanti, E., Filetti, P., Marchignoli, G., Ciociano, A., and Salvatore, R. 1995b. Strumenti evoluti per il supporto alla costruzione di banche dati legislative: estrazione automatica di riferimenti normativi, in *A.I.C.A.95 – Proceedings of the Annual Conference* , Chia (Cagliari, Italy) September.

Rama, D.V. and Padmini, Srinivasan, 1993. An investigation of content representation using text grammars, *ACM Transactions on Information Systems* **11**(1), 51–75.

Rau, L.F. and Jacobs, P.S. 1991. Creating segmented databases from free text for text retrieval, *Proccedings of the 14th International Conference on Research and Development in Information Retrieval*, SIGIR'91, ACM.

Salton, G. and McGill, M.J. 1984. *Introduction to Modern Information Retrieval*, McGraw-Hill.

Wright, G.H. von 1963. Norm and action: A logical enquiry. *International Library of Philosophy and Scientific Method*. Routledge & Kegan Paul, London.

XML 1998. *Extensible Markup Language (XML)*, 1.0 specification. W3C Recommendation, February