

A PUBLISHING SYSTEM TO EXTRACT AND REPRESENT THE KNOWLEDGE CONTENT OF SCIENTIFIC ARTICLES ON HEALTH SCIENCE IN MACHINE-PROCESSABLE FORMAT

Carlos Henrique Marcondes¹; Marília Alvarenga Rocha Mendonça¹; Luciana Reis Malheiros²; Leonardo Cruz da Costa³; Gabriela Veras De Moraes⁴.

¹ Department of Information Science

² Department of Physiology and Pharmacology

³ Department of Computer Science

⁴ Student, Biomedicine

Universidade Federal Fluminense

R. Miguel de Frias, 9, Icarai, Niterói - RJ, Brazil

e-mail: marcon@vm.uff.br

Abstract

Scientific articles published in electronic format are knowledge bases, especially in Medicine. An obstacle to semantic processing of this knowledge by computers is that in spite of their digital format, articles are in text format for human reading and processing. A model is proposed for electronic publishing scientific articles both in textual format and in machine “understandable” format, in ontology format. Software agents can process the content of an article published according to the model, thus enabling semantic retrieval, consistence checking and the identification of new discoveries. The model is described and initial steps toward the development of an authoring/publishing system which implements the model proposed are related.

Keywords: electronic publishing; scientific communication; knowledge representation; ontologies; e-Science.

1. Introduction

Biomedical research today is information driven. Large-scale databases with genomics, physiology, population genetics and imaging data are driving research at

increasing rates [1]. Their contents cannot be effectively processed to their full potential for research without the aid of computational tools. Information technology has speeded up electronic Web publishing and scientific communication, providing scholars with rapid access to recently published articles. However, scientific communication is still a slow social process which largely depends on discourse, text producing and reading/interpreting/inquiring these texts by scholars until the new knowledge is incorporated in the corpus of Science. The potential of information technology (IT) has been applied to bibliographic information systems as digital libraries, repositories and electronic journal systems to improve scientific communication, providing quick notification and immediate access to full-text scientific articles. But IT is not yet used to directly processing the knowledge embedded in the text of scientific articles. Current bibliographic information retrieval systems are based on sets of keywords to represent article content and the queries made to it. These keywords can only be related by Boolean connectors which lack semantic expressiveness. In the Semantic Web context [2], electronic publishing can be a cognitive tool, whose potential is far from being explored. Today electronic journals are still based on the paper print model. Electronic Web published articles are knowledge bases, but also for human reading.

Before the growth of the Web, what constituted approved scientific knowledge of humanity was fuzzy, it lacked formalization and was scattered across journal collections throughout libraries. Today there are two main barriers for large scale use of this knowledge: the amount of information available throughout the Web and the fact that knowledge is embedded in the text of scientific articles in an unstructured way, not adequate for program processing. Today different scientific communities are developing Web ontologies which formally record knowledge in different domains. In the near future, formal ontologies will be developed and recorded in machine-processable formats, containing knowledge in specific domains.

This research is pursuing a new paradigm in scientific electronic publishing. In the proposed model, each article, besides being published in textual format, is also published as an enhanced set of metadata which extends conventional bibliographic ones, aimed to represent both the reasoning process and the claims made by authors throughout the text of the article. These richer knowledge surrogates are represented as ontology instances in machine-processable format, which can be processed by software agents, retrieved and compared/mapped with the content of public ontologies available throughout the Web, especially in the field of Health Science, like the UMLS – Unified Medical Language System [3]. Articles published according to the proposed model may provide scholars with new tools for knowledge retrieval, identification and validation of new contributions to Science made by specific articles. The goal is to enhance Web electronic publishing embodying new facilities provided by the Semantic Web environment in order to create a pub-

lishing facility for e-Science environments [4].

We developed a content model for scientific articles as an ontology. The approach used for knowledge representation is based on the fact that scientific knowledge consists of propositions made in texts of articles, establishing relations between a scientific phenomenon and its characteristics or by establishing relations between different phenomena [5]. Relations in the texts of scientific articles appear as *questions*, *hypotheses* or *conclusions*. It is assumed that knowledge in the text of articles – scientific methodology elements such as the Problem, Hypothesis, Results and Conclusions – are all interrelated, constituting the semantic content and the reasoning process developed by the authors through which they communicate research results and new discoveries. With the support of a Web authoring/publishing tool claims made by the authors will be identified, extracted, marked up and recorded in a machine-processable format as ontology instances of the content model. This knowledge representation format can also be annotated and linked to public Web ontologies, enabling the establishment of formal relations between the text of the article as a knowledge base and Web ontologies which more and more represent the corpus of public knowledge in specific domains. Failure to establish these relations may be evidence of new discoveries, since they can indicate that the knowledge in the article is not yet represented in the ontology.

To enable such a publishing model the aim of this research is to develop an authoring/publishing software tool that offers researchers/authors an interactive web environment, which through an interactive dialog and using natural language processing and text extraction techniques, identifies and extracts the semantic elements comprising the knowledge content of the article being published. The knowledge content surrogates of articles holding the claims made by the authors and details of their experiment are then represented and recorded as ontology instances.

How to extract these knowledge surrogates from the text of articles? Which questions must be proposed by the authoring/publishing tool to an author? How to process his/her answers? How to map them in the developed ontology model?

This is a report of a research in progress. In its first stage we developed a rich and complex content model to be represented in a machine readable format as an ontology [6]. In the current stage the aim is to develop the authoring/publishing system to extract knowledge from the text of articles and to represent it in a structured format feasible for machine processing.

To extract structured information from text is a challenge. But even simple and incremental approaches to the content model here proposed may have important impacts on knowledge management and retrieval throughout the Web, providing scientists with tools which enable them to trace the development of an idea or hypothesis, to identify who is working on a specific problem or which answers have

been given to a question.

The rest of the paper is organized as follows: First, the material and methods that were used are presented. The results obtained so far are presented, outlining the proposed content model and the choices made in developing procedures to extract knowledge content from texts of the articles. The next Section discusses the results, presenting features, potentialities and challenges of the model. Finally, the last section presents the conclusions and outlines the contribution of this research.

2. Material and Methods

A conceptual content model was first developed by the analysis of 75 articles in the field of Health Science. Articles in Health Science were used in the test due to their highly structure format - IMRAD: Introduction, Material, Results and Discussion [6]. Protegé software was used to implement the content model as an ontology in OWL. A tentative interactive dialog with the authors has been also developed with the aim of selecting pieces of text to be automatically processed, to confirm the results of natural language processing of the text of an article or of answers provided by the authors. We have performed interviews with the authors of Health Science articles to decide the best dialog strategies to implement the future system. The MetaMap Transfer program [7], from the National Library of Medicine, USA, is used in order to syntactically process text utterances from questions posed to the authors and from sections of articles as title, abstract and conclusions and to map identified phrases of concepts of UMLS Metathesaurus [8].

3. Results

Relations are essential to the proposed knowledge representation schema. Its relations are expressed by three elements: two relata and a type of relation. The two relata may be two different phenomena or a phenomenon and one or more of its characteristics.

Semantic elements in scientific articles are structured to form an ontology in the sense used in knowledge engineering [9]. Article surrogates in machine-processable format are instances of this ontology. Articles surrogates have links between their semantic elements and the corresponding bibliographic metadata and full-text in a digital library or electronic journal publishing system.

The semantic elements which comprise the Ontology for Knowledge in Articles – the OKA model – the classes are the following:

- the **problem** the article is addressing, which leads to one or more **questions** the article is trying to answer;
- the **hypothesis**, in which the author states a **relation** between phenomena or, in the case of Experimental-exploratory articles, between a phenomenon and its characteristics. Hypotheses are important for knowledge management of the content of the articles as they have the form of relations between phenomena, and thus can be used by software agents to make inferences. An original hypothesis, formulated by the author of the current article is named **new hypothesis**; a hypothesis formulated in a previous article by an author who is different from that of the current article's author is named **previous hypothesis**. A **hypothesis**, as a relation, has two **arguments** related by a **type_of_relation**:
 - an **antecedent**,
 - a **type_of_relation** (holding the semantic of the relation in a domain, for ex., in Health Sciences), and
 - the **consequent**; **antecedent** and **consequent** may be two different phenomena or a phenomenon and its characteristics;
- a possible empirically controlled **experiment** with the aim of observing the phenomenon described, specifics of experimental articles, divided into
 - **results** - tables, figures, numeric data, reporting the observations made;
 - a **validation_grade** of the results obtained specifying if the results deny the hypothesis, or not yet confirm the hypothesis, or partially confirm the hypothesis or if they confirm it;
- **measure** used;
- a specific **context** where the empirical observations take place, subdivided into:
 - **environment** - a hospital, a day-care center, a high school, a geographical **place** where the empirical observations take place,
 - **time** that the empirical observations occurs, a specific **group** - pregnant women, premature babies, mice - in which the phenomenon occurs,
 - **conclusion** - a set of propositions made by the author as a result of his/her findings;
- a **conclusion**.

Not all those elements are present in all articles.

Articles differ in the way they are built around previously stated hypotheses: those stated by the authors other than the one of the current article; or new, original hypotheses – those stated by the author of the current article. Articles may also differ by the existence of a documented experiment or just theoretical considerations comparing previously stated hypotheses. We found four patterns of reason-

ing in the analyzed articles: *theoretical articles*, which employ abductive reasoning and *experimental articles*, which may be just *exploratory* or employ *inductive* or *deductive* reasoning. This schema is the result of Kintsh & Van Dijk [10], Hutchins [11], Gross [12], Kando [13], [14], Thagard [15] and Klahr & Simon [16] proposals as well as the result of the analysis of 75 articles to find reasoning patterns as previously described.

Theoretical-abductive articles analyze different, previous hypotheses, showing their faults and limitations and proposing a new hypothesis.

Experimental-inductive articles propose a hypothesis and develop experiments to test and validate it. In experimental-inductive articles, a **conclusion** may be mainly one of these alternatives: either it corroborates the hypothesis, refuses it, or partially corroborates the hypothesis. However, in some cases, the Conclusion is not one of the former; it just reports intermediate, not conclusive results toward the hypothesis corroboration.

Experimental-deductive articles use a hypothesis proposed by other researchers, cited by the author of the articles and apply it to a slightly different context.

Experimental-exploratory articles are not usually hypothesis driven; their objective is to acquire knowledge about a poorly understood scientific phenomenon by performing an **experiment**.

The different reasoning procedures previously discussed can be formalized in the Ontology for Knowledge in Articles as illustrated in Figure 1, having the following Classes and Properties:

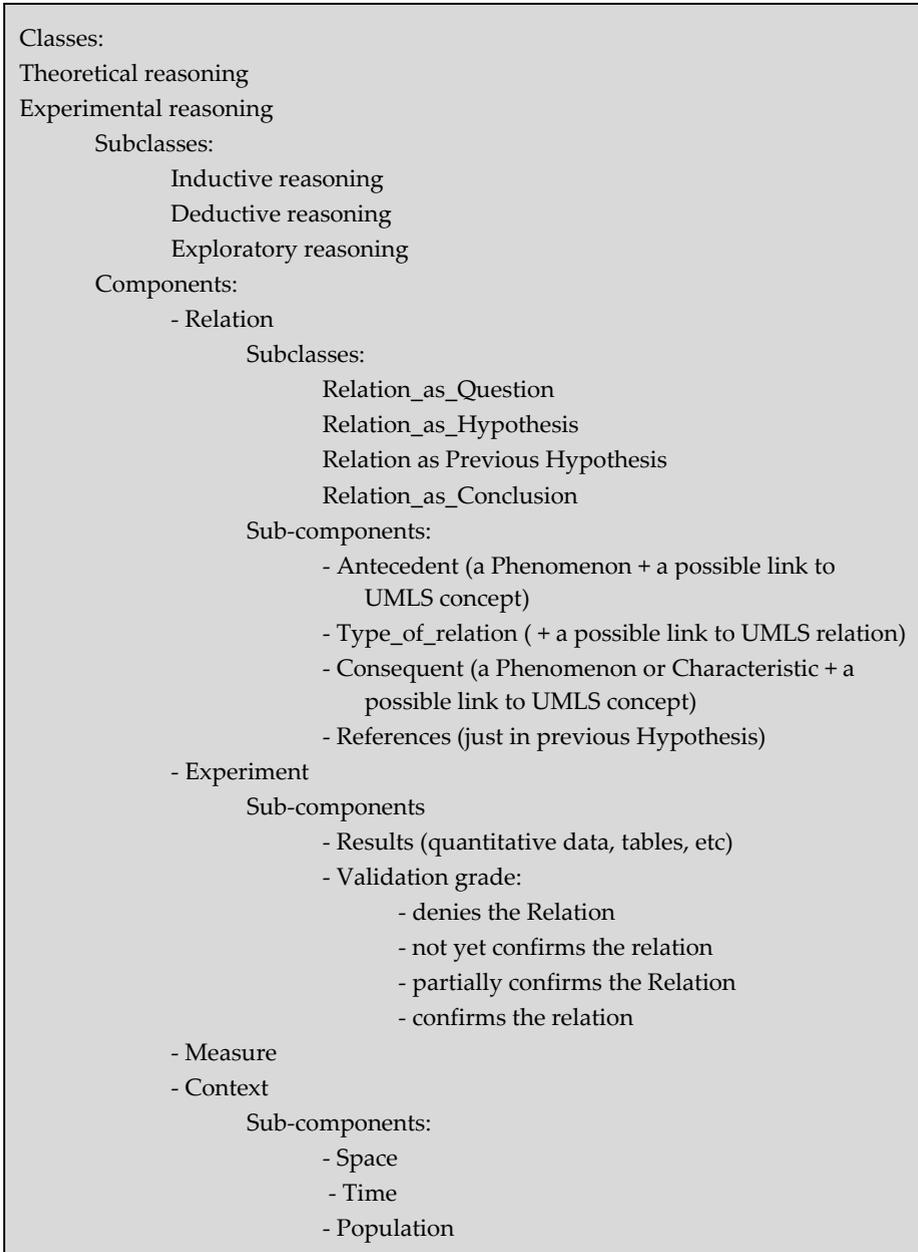


Fig.1. Model of the Ontology for knowledge in scientific Articles

The implementation of such a model poses twofold challenges: the first is how to extract relations from the text of an article, a key feature of the model proposed; the second is to use article surrogates in compliance to the model in semantic retrieval and knowledge management tasks.

We propose to address the first challenge by involving authors with tasks as disambiguating natural language expressions. Today, researchers are accustomed to self-publishing, self-describing their papers when submitting them to a digital library, to a conference, to a digital repository or to a journal system. They are also accustomed to writing structured abstracts [17], mainly in the field of health sciences. We envisage an authoring/publishing system that offers researchers/authors an interactive web environment, which through a rich dialog and using natural language processing and text extraction techniques, interactively identifies and extracts the semantic elements of the article being published. The pathway which seems most feasible to reaching this objective is to provide authors with an interactive interface which enables them to validate the automatic natural language processing made by the system. We intend to combine natural language processing of the article text with the processing of the text of answers provided by authors to questions proposed by the authoring/publishing system. The interface asks authors to enter the Problem the article addresses and the Question it is trying to answer, the working Hypothesis, the claims synthesising the article Conclusions. These answers are inputs to natural language processing. Some elements of the OKA model are either explicit sections in the text of articles or can be directly asked from authors, for example, whether the article is theoretical or experimental, whether the results confirms or deny the hypotheses, whether the article is based on hypothesis of other authors or if its hypothesis is original, etc.

Relations are the core of the representational schema here proposed. The identification of relation is harder. A relation is generally posed by a verb phrase or a verb noun. The challenge is to identify the model semantic elements, especially relations, using natural language processing techniques in pieces of the text as the problem, the title, the abstract, the objective, the hypothesis, the questions posed by authors and the conclusion. We believe that natural language processing will be necessary in very short texts; what makes it more effective and error free, as for example, in the processing of answers provided by the authors to questions made by the system ("Which is your main conclusion in this article?") or in explicitly in article titles.

The results of interviews with authors of articles we analysed indicate that asking them the conclusion posed in the article generally matches manual analysis results. The titles of articles on the contrary, frequently report methods used, not the conclusions, or use indirect speech or noun verbal phrases. When asked for the

conclusions of their articles, authors generally present clear, concise and direct statements.

The linguistic analysis is performed in the following steps:

- a - Text in the article's Introduction is processed in order to identify what the article's Objective would be; the Objective identified by the system is displayed to the authors and he/she is asked to validate the Objective; if he/she does not agree with the identified Objective the system asks the author to enter as the Objective.
- b - Text in articles title, abstract, question, objective and conclusions were analyzed by MetaMap. MetaMap program parses natural language utterances into components as phrases and assigns syntactic categories to these such as VERB_PHRASE (candidates do be relations), NOUN_PHRASE (candidate to be phenomena) or PREP_PHRASE (candidate to be context information).
- c - NOUN_PHRASEs found are weighted according to their frequency and their presence in questions, title, objective, abstract, hypothesis or conclusions;
- d - The two first raked not adjacent NOUN_PHRASEs are the candidates to be Phenomenon which will be mapped to the two relata of a relation;
- e - The system then looks for a VERB-PHRASE occurring in the text between the two NOUN_PHRASEs previously identified as a candidate to be the relation. A relation dictionary is developed on the base of the 53 relations (and their synonyms found in the WorldNet Dictionary) which comprise UMLS Semantic Network [18].

A semantic grammar was developed to process the syntactic categories instances found in the previous phase into elements of the content model. This grammar comprises elements as Phenomena, Relation and Context_indicator which precede text with context or methodological information. The semantic grammar recognizes the following pattern of semantic elements in text utterances:

Phenomenon, Relation, Phenomenon,
1{(Context_indicator + context information)}n.

- f - These linguistic level elements previously identified are then mapped to the content model elements, the Phenomenon to Antecedent, the Relation to Type_of_relation, the second Phenomenon to the Consequent, context information and methodological information to the Context. After the semantic content model elements obtained are mapped to concepts in UMLS. The MetaMap program is used in order to map semantic content elements identified to concepts of UMLS Metathesaurus. At this phase we intend to ask the author for validation both of the relation identified and of its mapping to

UMLS concepts. Authors will browse UMLS Metathesaurus and will be asked to validate the mapping of the semantic model components identified in the UMLS concepts, thus annotating the article ontology instance with these concepts and recording a possible link to the corresponding concepts in this ontology.

Throughout this analysis process the knowledge content of the article can be identified, represented as an ontology instance and recorded in machine-processable format, comprising the knowledge representation surrogate of the article content.

4. Discussion

The second challenge may be addressed by using Semantic Web standards as OWL - the Ontology Web language to represent the content of articles as ontology instances. This option enables a set of surrogates of articles, thus represented to be queried using SPARQL query language and data access protocol for the Semantic Web [19]. SPARQL enables queries to multiple data sources – journal systems, digital libraries, repositories – and integration of results to perform inferences and knowledge management tasks.

The proposed knowledge model and its coding in OWL will enable queries as such:

- What other articles have hypotheses suggesting HPV as the cause of cervical neoplasias in women? Which of those have proven hypotheses?
- Which articles have hypotheses suggesting other causes to cervical neoplasias different from HPV in women?
- Which articles have hypotheses suggesting HPV as the cause of cervical neoplasias in populations different from women?
- Which articles have hypotheses suggesting HPV as the cause of other pathologies different from neoplasias?
- Which experimental-inductive articles propose (antecedent ?) causes (type_of_relation) to cellular senescence (consequent) which are not-mapped to UMLS concepts?

Is there any confirmation of the hypothesis that “Several aspects of both the structural and dynamic properties of telomeres led to the proposal that telomere replication involves nontemplate addition of telomeric repeats onto the ends of chromosomes”? [20].

- Who and when first maintained that “the RNA component of telomerase may be directly involved in recognizing the unique three-dimensional structure of the G-rich telomeric oligonucleotide primers” (GREIDER, 1987)?

It is very difficult to collect natural language utterances to use in testing our prototype systems because we do not have a real working authoring/publishing environment. We have interviewed authors to collect information about previously published articles. These interviews try to simulate the dialog to be performed by an interactive authoring/publishing system. Notwithstanding these difficulties, we are getting increasing evidence through the interviews performed with authors that the conclusion of an article, more than the titles or abstract, synthesizes its discoveries and its contribution to science.

5. Conclusions

Electronic publishing of scientific articles both as text and as machine-processable knowledge bases as outlined seems to be a step ahead of the conventional paper print model of today's electronic journals and a valid research objective. It points toward a whole integrated e-science environment where scientists can take advantage of the large scale processing of knowledge content of scientific articles and their comparison to Web public ontologies and other electronic resources.

Notes and References

- [1] STEIN, L.D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews Genetic*, 9, Sept. 2008, p. 678-688.
- [2] BERNERS-LEE, T., HENDLER, J. and LASSILA, O. The semantic web. *Scientific American*, May 2001. Available at <<http://www.scian.com/2001/0501issue/0501berners-lee.html>>.
- [3] UMLS - Unified Medical Language System, the UMLS Semantic Network (2005). Available at: www.nlm.nih.gov/pubs/factsheet/umls.html (accessed 5 Mar. 2006).
- [4] DE ROURE, D., JENNINGS, N. and SHADBOLD, N. *Research agenda for the Semantic Grid: a future e-Science infrastructure*, Report commissioned for EPSRC/DTI Core e-Science Programme, 2001.
- [5] MILLER, D. L. Explanation Versus Description, *Philosophical Review* 56(3), 1947, p. 306-312.
- [6] MARCONDES, C.H., MENDONÇA, M.A.R., MALHEIROS, COSTA, L.C. da; SANTOS, T.C.P. and PEREIRA, L.G. Representing and coding the knowledge embedded in texts of Health Science Web published articles. In: Chan, Leslie; Marten, Bob, ed. ICCC EIPub - INTERNATIONAL CONFERENCE ON ELECTRONIC PUBLISHING, Vienna, Austria, 2007, 11, *Proceedings...* Vien,

- Austria, 2007. Available at <http://elpub.scix.net> (July 2007).
- [6] International Committee of Medical Journals Editors, 2003. Available at: www.icmje.org (Jul. 2005).
- [7] *The MetaMap Transfer program*. Available at <http://mmtx.nlm.nih.gov/> (March 2009).
- [8] *UMLS Metathesaurus*. Available at <http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html> (March 2009).
- [9] SOWA, J. *Knowledge Representation: logical, philosophical and computational foundations*. Brooks/Cole : Pacific Grove, 2000.
- [10] KINTSH, W., VAN DIJK, T.A. Towards a model of text comprehension and production. *Psychological Review* 84(5), 1972, p. 363-393.
- [11] HUTCHINS, J. On the structure of scientific texts. In: *Proceedings of the 5th. UEA Papers in Linguistics*, Norwich, Norwich, UK : University of East Anglia, 1977. p.18-39. Available at: <http://ourworld.compuserve.com/homepages/wjhutchins/UEAP/L-1977.pdf> (March 2006).
- [12] GROSS, A.G. *The Rhetoric of Science*. Cambridge, Massachusetts; London: Harvard University Press, 1990.
- [13] KANDO, N. Text-level structure of research papers: implications for text-based information processing systems. In: J. Furner and D.J. Harper (eds.), *Information Retrieval Research: Proceedings of the 19th BCS-IRSG Colloquium on IR Research*, Aberdeen, 1997. Aberdeen, Scotland: Springer-Verlag, 1997.
- [14] KANDO, N. Text structure analysis as a tool to make retrieved documents usable. In: *Proceedings of the 4th International Workshop on Information Retrieval with Asian Language*, Taipei, 1999. Taipei, Taiwan : Academia Sinica, 1999.
- [15] THAGARD, P. *Computational Philosophy of Science*. Cambridge, MA: The MIT Press, 1993.
- [16] KLAHR, D. and SIMON, H.A. Studies of scientific discovery: complementary approaches and convergent findings, *Psychological Bulletin* 125(5) (1999) 524-543.
- [17] BAYLEY, L. and ELDREDGE, J. The structured abstract: an essential tool for researchers. *Hypothesis* 3 Spring, 17(1), 2003, p.11-13. Available at: gain.mercer.edu/mla/research/hypothesis.html (Jul. 2007).
- [18] UMLS Semantic Network. 2006. Available at <http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html> (March 24 2009).
- [19] SPARQL Query Language for RDF. 2008. Available at <http://www.w3.org/TR/rdf-sparql-query/> (March 2009).
- [20] SHAMPAY, J., SZOSTAK, J.W. and BLACKBURN, E.H. DNA sequences of telomeres maintained in yeast. *Nature* 310, 1984, p. 154-157.

June 2009
Printed on demand
by "*Nuova Cultura*"
www.nuovacultura.it

Book orders: ordini@nuovacultura.it