Received 12 January 2007
Revised 16 January 2007
Accepted 16 January 2007

Social metadata for the impact factor

Nikolaos Korfiatis
Department of Informatics, Copenhagen Business School, Copenhagen, Denmark, and
Marios Poulos and George Bokos
Department of Archives and Library Sciences, Ionian University, Corfu, Greece

Abstract

Purpose – The purpose of this research is to address the need for a definition of metadata descriptors for use in enhancing the accuracy of bibliometric instruments of scholarly evaluation, such as the impact factor.

Design/methodology/approach – A semantic vocabulary – COAP – is constructed, deployed on top of the Resource Description Framework (RDF), by extending the Friend-of-a-Friend (FOAF) schema.

Findings – An extension of the FOAF vocabulary is considered as the ability to describe a publication record such as this paper in terms of scholar contributions and participations. In order to achieve that, the FOAF vocabulary is extended.

Practical implications – The application of this semantic vocabulary could be used as a way of enhancing the accuracy of source data for bibliometric evaluation instruments.

Originality/value – The paper discusses how metadata descriptors can contribute to the improvement of already established scholar evaluation instruments such as the impact factor. It will be of use in the development of digital libraries.

Keywords Social networks, Digital libraries, Semantics

Paper type Research paper

Introduction

Since the appearance of the early forms of scholarly publishing, the communication process that characterizes the distribution of scientific knowledge among scholars, continues to stay untouched (Odlyzko, 2002; Resh, 1998). Scholars communicate research results by publishing papers on journals and conferences or by other types of scientific artifacts (raw data, reports on research prototypes, software), either on their own either with collaborators, with whom they establish a communication relationship. Furthermore as a result of scholar communication, scholarly works such as publications, are characterized by the product of an authoritative entity or a set of entities affiliated with an organization or an institution. With this authoritativeness or co-location collaboration among scholars emerges in the form of co-authorship or citation of previous work.

Concerning the communication and collaboration capabilities, with the support of information technology the realization of infrastructures that support the discussion of ideas and communication among scholars has evolved to what is known today as electronic scholar communication (Resh, 1998) making scholar collaboration easier than ever before. In particular the development of the main communication infrastructures that characterize to a significant degree today's work practices, such as the World Wide Web, has been motivated from communication needs of the scientific community (Berners-Lee and Fischetti, 1999). As a result there is a consensus that scientific production has

The Electronic Library
Vol. 25 No. 2, 2007
pp. 166-175

0264-0473
DOI 10.1108/02640470710741304

increased in terms of publications and material available for reference. However the critical question remains: how to evaluate the quality of those productions and to an extend the scientific output of an individuals work? With the development of the WWW a large amount of scholar communication data (co-authorship, citations) became available to collect and examine making easier to utilize this data towards the development of scholar evaluation instruments. As has been seen in the case of the Impact Factor (IF), those developments have advanced the role that bibliometrics or informetrics (Egghe and Rousseau, 1990) can play on the improvement of the quality of scientific work.

In the literature there are several cases whereas the field of bibliometrics has considered the exploration of co-citation networks and social networks based on co-authorship relations as a source of evaluation (Newman, 2001). These that can enhance the impact factor and other instruments of evaluation, however, little has been done to improve the source data that can be used on these studies. In particular one of the main critical claims over the use of evaluation instruments on scholarly work such as citation indices is the poor quality of the data available that can be a trustworthy source (Hicks, 1999; Moed, 2002). The development of citation crawlers such as Citeseer (http://citeseer.ist.psu.edu/) and recently Google Scholar (http://scholar.google.com) has provided a step towards that direction however natural language processing issues as well as the ambiguity under which a citation is provided, e.g. using different versions of the same paper had until recently a negative impact on their development.

On the other hand in the field of digital libraries and in particular the research branch dealing with creation and definition of metadata descriptors has not put much effort on providing metadata that describe co-authorship data in a more detailed degree than the abstract cataloging metadata schemas. Such an example is the definition of the academic metadata format (AMF) which is developed in the context of Dublin Core (Krichel and Warner, 2001).

Therefore the scope of this paper is twofold. First it addresses the issue of providing effective metadata descriptors in order to cover the process of scholar communication and collaboration and in particular discussing the way how those descriptors could contribute to the improvement of already established scholar evaluation instruments such as the impact factor (Garfield, 1972). We do this by introducing the COAP semantic vocabulary, an extension to the widely deployed FOAF vocabulary (Brickley and Miller, 2006) which is based on the Resource Description Framework (RDF) (Lassila and Swick, 1999). In respect of the second objective, we describe issues on binding those metadata to evaluation metrics based on egocentric networks that can provide both an indication of the activity and the prominence of a scholar.

Following the above objectives, this paper is structured as follows. Section 2 provides a review of the impact factor and related bibliometric concepts from a metadata perspective. Section 3 provides an overview of the FOAF vocabulary and the contributed extension. Section 4 discusses binding of the vocabulary properties to sociometric ties and possible metrics that can be used such as the centrality. We conclude this paper in section 5 by listing future extensions and some practical considerations regarding the applicability of the proposed extension.

Metadata requirements for the impact factor

The impact factor (Garfield, 1972) has been undoubtedly the main instrument for evaluation of the importance of scientific publications which transpose some of their

credibility to the contributing authors. In particular most authors consider as an important step on becoming prominent in their field/scientific community to publish in a set of highly prestigious journals measured by their impact factor. Furthermore the way impact factor is been calculated raises some criticism over its trustworthiness as an indicative measure of academic prestige towards the evaluation of research works published in journals with high impact factor. Nonetheless the usage of impact factor has often been a point of criticism whether it should be used as an instrument for evaluation or not (Seglen, 1997; Coleman, 1999).

In particular for a period of time t the impact factor for a journal is calculated as the fraction of the number of citations given to that journal divided with the number of articles published, with both measured in the period t. To this the first obvious argument is made from the side of citation extraction from the databases, where the citations are often not clearly expressed or self citations are not omitted. Another particular issue is the bias of the database towards limited resources in the scientific field. For instance in fields so wide such as physics, a sociology oriented paper published in a physics journal might attribute citations to the journal which in general are not committed by publications on its own research topic.

Scholars have considered the above issues and have produced the development of more trustworthy measures that rely on several factors that characterize the academic output apart from the citations *per se*. There are several directions in the area such as approaches on calibrating the impact factor along with the co-authorship/citation network of each author. (Krichel and Warner, 2001), Electronic publication models that rely on readers and their competencies in order to attribute trustworthiness on a particular article (Mizzaro, 2003) and finally totally alternative methods of evaluation such as the H-Index (Hirsch, 2005). The last one considers data about publications and citations as an important point of departure towards the development of a probabilistic model of scholar evaluation which provides a point of reference in an optimum scale.

For our research the way the impact factor is measured and thus a scholars' prominence is evaluated, contributes an important remark to metadata requirements. In particular the definition of a measure of visibility/prominence in the scientific area continues to suffer by incomplete data. However with the support of digital libraries and publication repositories such as Arxiv (www.arxiv.org) a considerable amount of metadata could be produced and used in the development of an evaluation metric.

The FOAF vocabulary

We refer to the definition of Social metadata can be attributed as metadata that can be used as descriptors of social relations and in particular properties of those relations such as strength and level of relation (Mika and Gangemi, 2004). Probably the most popular social metadata schema is the Friend-of-a-Friend vocabulary commonly known as FOAF (www.foaf-project.org) (Brickley and Miller, 2006). The Friend-of-a-Friend vocabulary is an expressive vocabulary set which syntax is based on RDF syntax (Klyne and Carroll, 2004) that is gaining popularity nowadays as it is used to express the connections between social entities in the web along with their hypertextual properties such as their homepages or the emails. Table I describes some of the FOAF elements and their type of relational tie.

When considering metadata as a descriptor we need to consider two perspectives respectively: the vocabulary and the syntax. To support our approach we make use

of RDF syntax rather than the syntax provided by an XML model for two basic reasons: First reason is that RDF provides a richer way on expressing descriptions in a machine processable way by avoiding issues such as polysemy and synonymy usually met in the tree structure of XML documents. More specifically in the case of linking authors with a co-authorship relation, the RDF structure of the document permits an interoperable matching of the resources due to its graph nature which cannot be done in tree based XML structure. Furthermore RDF allows for the incomplete characterization of resources whereas in the case of constructing the profile of an author, basic inference can be done without committing to an RDF schema.

Extending the FOAF vocabulary

Academic work on extending the FOAF vocabulary on the way towards providing meaningful descriptors for various cases of social relations has been undertaken in academia such as the extension with the trust ontology (Golbeck *et al.*, 2003) to represent levels of trust between FOAF:person. Various non-academic purpose extensions of FOAF have also been provided and discussed, e.g. resume management, reviews annotation etc. The SchemaWeb portal (www.schemaweb.info/) contains most of the academic and non-academic extensions to FOAF currently published in the FOAF namespace.

What we consider as an extension of the FOAF vocabulary is the ability to describe a publication record such as this paper in the term of scholar contributions and participations (Table II). In order to achieve that, we extend the FOAF vocabulary as follows. We consider a class on the range of FOAF:person which we define as

Vocabulary element	Description	Type of relational tie
foaf:knows	Links foaf:persons	Direct
foaf:member	Provides affiliation/membership (relates an entity with a social group)	Indirect
foaf:maker	Indicates authorship (relates an information resource with its creator)	Indirect
foaf:based_near	Indicates a spatial affiliation of a social entity	Indirect
foaf:currentproject	Indicates a temporal affiliation of a social entity with a project or an activity	Indirect

Table I.
Description of elements from the FOAF vocabulary along with their type of relational tie

Vocabulary element	Description
coap:author	Represents an author subclass of FOAF person
coap:coauthor	Indicates a co-authorship relations between coap:author
coap:publication	Encapsulates the publication details
coap:affiliation	Describes the affiliation, temporary or permanent, which the coap:author had when published the paper

Table II.
The basic elements of the COAP semantic vocabulary

< coap:author > . We then define an abstract property < coap:coauthor > on the range of < foaf:knows > property available in the FOAF vocabulary. By doing that we can define a logical sequence as a triple statement (Figure 1). For example Nikos coauthors a paper with Marios, with Nikos and Marios having been instantiated as < coap:author > . The coauthor property denotes whenever this person has co-authored a resource with someone else. As we will discuss later the < coap:coauthorof > property is important for the construction of egocentric networks of authors with their collaborators. Central to the extension is the definition of the < coap:publication > Apart from the co-authorship we consider other cases of

```
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:coap="http://www.korfiatis.info/papers/tel/coap/">
    >
<foaf:Document rdf:about="">
 <rdfs:label>Demonstration of the COAP vocabulary</rdfs:label>
 <rdfs:comment>Document with Friend-of-a-Friend description of eikeon.</rdfs:comment>
</foaf:Document>
<foaf:Person>
 <coap:Author>
        <foaf:name>Nikolaos Korfiatis</foaf:name>
        <foaf:mbox rdf:resource="mailto:nk.inf@cbs.dk"></foaf:mbox>
        <foaf:homepage rdf:resource="http://www.cbs.dk/staff/nikos"></foaf:homepage>
        <coap:currentAffiliation>Copenhagen Business School</coap:currentAffiliation>
 </coap:Author>
</foaf:Person>
<coap:publication rdf:nodeID="coap2006">
<dc:title>Social Metadata for the Impact Factor</dc:title>
<dc:year>2006</dc:year>
</coap:affiliation rdf:resource="#currentAffiliation">
</coap:publication>
<coap:publication rdf:nodeID="MSThesis">
<dc:title>The Opinion Evaluation Network: Ranking Imprecise Social Interactions</dc:title>
<dc:year>2005</dc:year>
<coap:affiliation>Royal Institute of Technology, Sweden</coap:affiliation>
</coap:publication>

<coap:coAuthor>
<coap:author>
<foaf:person>Marios Poulos</foaf:person>
</coap:author>
</coap:contribution rdf:nodeID="coap2006">
</coap:coAuthor>

<coap:coAuthor>
<coap:author><foaf:person>George Bokos></foaf:person></coap:author>
</coap:contribution rdf:nodeID="coap2006">
</coap:coAuthor>
</rdf:RDF>
```

Figure 1.
Description of the co-authorship relations of that paper using the coap schema

informal connections such as the affiliation which we define as temporal or current. Temporal is an affiliation that is not existent outside the time-space of the publications that we consider as an evaluation source.

Although the set of classes and properties in the existing version of the FOAF vocabulary contains a considerable set of elements that could be used to describe the publication record of an individual we considered to subclass those elements that correspond to the property or class we need to describe to a new subset which is deployed for the selected purpose.

Previous work in the field has addressed the scholar communication as a descriptive element in the web resource without providing any reference to its context. For instance as aforementioned the Academic Metadata Format (AMF) definition extends on the Dublin Core Metadata structure to provide a description of coauthorship as dc:contributor. However this description inherits the disadvantages of tree based XML based descriptions which is bounded to the content of the examined profile.

Development of metrics using the COAP vocabulary

In order to bind our semantic vocabulary in a sociometric analysis which can be later used to calibrate the impact factor we need first to consider the type of variables that are needed. This can be seen in two dimensions: the first class of variables elaborates on attributes/properties of the referencing entity – the ego. The second class of variables deals with the relations of the ego with the rest of the egos in the network – seen as alters. Apart from direct ties there can be indirect ties between two authors by the use of indirect properties such as the coap:affiliation which can link two authors who haven't had a co-authorship relation before. However the type of the relation needs to be taken into account and in particular the strength arising from the formal or informal connection.

Several bibliometric studies on digital libraries have considered the socio-structural properties of the authors in their co-authorship network as a factor that can measure their scientific output (Liu *et al.*, 2005). Social Network Analysis (Scott, 2000; Wasserman *et al.*, 1994) has been used in several bibliometric studies in order to explain mostly how a scientific community evolves in terms of collaboration and how influential members of this community shape the research field discussed.

Figure 2 depicts a visualization of a subset of an ego based co-authorship network from authors obtained from the DBLP Electronic Library. Author names are substituted with numbers. Edges indicate a coap:coauthor relational tie depicted in author profiles. Apparently the most well known informal type of egocentric network among scholars is the Erdös number (www.oakland.edu/enp/) used in the community of mathematicians and theoretical computer scientists. The Erdös number of a scholar provides a descriptor of the distance of the alter from the ego which in that case the ego is Paul Erdös. Considering the impact and the productivity of Erdös' work in the field of mathematics, one who has the lowest Erdös number (closest distance from the ego) is considered as an important alter to the ego thus is sharing some of the ego's credibility through this connection. In a more macroscopic way if we consider the case of affiliation networks a scholar affiliated with a renowned institution attributes some credibility to the paper which can be a factor that influences the citation by future readers. However there are many cases whereas that this kind of association is not

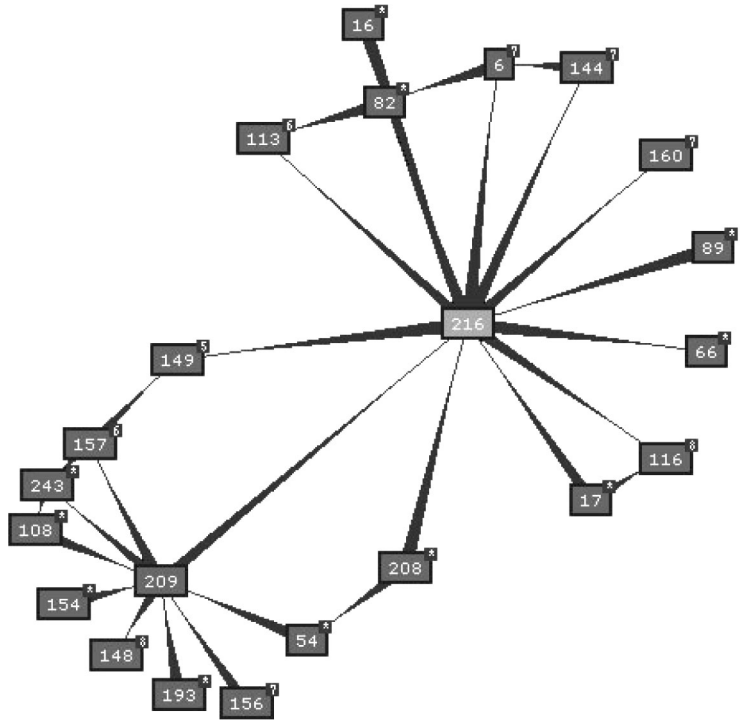

Figure 2.
Visualization of an ego based co-authorship network

reciprocal. This implies that a prestigious affiliation of the author doesn't guarantee a prestigious research profile.

The formalization of the egocentric network of an author follows the same formalization from Social Networks by using a special case of graph where the reference point is provided by a starting node referenced as the "ego". We consider a network $G := (V, E)$ with V_i being the ego and $N = |V| - V_i$ the set of alters. Depending on their connection with the ego the alter can be a first-order or in general λ-order where lambda represents the longest path between any alter node $n_i \in N$ and the ego node.

Having extracted the network relations from the coap:coauthor or the coap:affiliation properties of an author's profile within the context of a scientific community we can consider metrics that can evaluate the prominence of an author throughout his/her egocentric network. Such metric of prominence can be for instance the centrality based on the number of connections (degree centrality). In our case considering the ego node V_i the degree centrality can be calculated as follows (Marsden, 2002):

$$C_D(p_i) = \sum_{k=1}^{N} a(p_i, p_k)$$

Degree centrality in that case is a direct representation of prominence based on the amount of the coap:coauthor connections that this ego has – the number of alters connected to him/her.

Applications and future work

Applications of the coap vocabulary can be considered the evaluation of already established metrics for calibration of citation based rankings with sociometric based rankings such as visibility/status and prominence/centrality in a social network. For example Garcıa and Sicilia (2005) have extended the concept of social and citation based relevance in which the proposed semantic vocabulary could contribute by providing a ground for the construction of social relevance upon coupling on co-citation networks. Another aspect of that direction of research is the construction of informetric models of scientific collaboration whereas the development of COAP could provide a better ground for data sensitive processes such as the development of an informetric model (Egghe and Rousseau, 1990).

Another approach on the usage of coap could be the development of mechanisms for estimation of conflict of interest. In particular recent work by (Aleman-Meza *et al.*, 2006)has used the DBLP dataset and social metadata of the author profiles to model conflict of interest (COI) in the publication process. An ego can be influential about the review of a publication in case there is an alter within a very close distance degree which participates in the review process. Determination of conflict of interest directs to a future development of coap with the inclusion of properties such as events (e.g. common conferences) or membership in program committees and editorial boards.

However, in our work there are open issues that highlight possible directions for extending our semantic vocabulary. In particular:

- *The role or reciprocity in the co-authorship ties attributed by the metadata*. Status and prominence is a non reciprocal property of a particular entity. As it happens in a sociological setting a person might be popular but this fact doesn't necessarily assert that he/she will be prominent. If we consider it in the case of an author, his/her scientific production in a certain field, indicated by the coap:publication.
- *Quantification of the relational tie beyond the Boolean space and representation in the schema*. When we consider the representation of a tie in a co-authorship network we derive that representation from the coap:coauthor fragment in the FOAF profile by using a Boolean indication, e.g. present or not-present. However as it happens in a social context of a relation of collaboration vary on different degrees of strength. E.g. in the case of a supervisor and a PhD student and/or two colleagues of the same academic ranking working together. On the basis of this relations

Extension and realization of metadata schemas for describing scholarly communication and activity is a step towards improving the quality and trustworthiness of scientific evaluation instruments. The proposed extension to FOAF contributes to that direction.

References

Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A.P., Budak Arpinar, I., Joshi, A. and Finin, T. (2006), "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection", *WWW '06: Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland*, pp. 407-16.

Berners-Lee, T. and Fischetti, M. (1999), *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*, Harper, San Francisco, CA.

Brickley, D. and Miller, L. (2006), *FOAF Vocabulary Specification*, available at: http://xmlns.com/foaf/0.1/ (accessed July 2006).

Coleman, R. (1999), "Impact factors: use and abuse in biomedical research", *The Anatomical Record*, Vol. 257 No. 2, pp. 54-7.

Egghe, L. and Rousseau, R. (1990), *Introduction to Informetrics: Quantitative Methods in Library*, Elsevier Science Publishers, Amsterdam.

Garcia, E. and Sicilia, M.A. (2005), "Filtering information with imprecise social criteria: a FOAF-based backlink model", *Proceedings of the 4th Conference of the European Society For Fuzzy Logic and Technology (EUSLAT), Barcelona, Spain*.

Garfield, E. (1972), "Citation analysis as a tool in journal evaluation", *Science*, No. 178, pp. 471-9.

Golbeck, J., Parsia, B. and Hendler, J. (2003), "Trust networks on the semantic Web", *Proceedings of Cooperative Information Agents, Helsinki, Finland*.

Hicks, D. (1999), "The difficulty of achieving full coverage of international social science literature and the bibliometric consequences", *Scientometrics*, Vol. 44 No. 2, pp. 193-215.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences*, Vol. 102 No. 46.

Klyne, G. and Carroll, J.J. (2004), "Resource Description Framework (RDF) concepts and abstract syntax", *W3C Recommendation*.

Krichel, T. and Warner, S.M. (2001), "A metadata framework to support scholarly communication", paper presented at International Conference on Dublin Core and Metadata Applications, Japanese National Institute of Informatics, Tokyo.

Lassila, O. and Swick, R.R. (1999), "Resource Description Framework (RDF) model and syntax specification", *W3C Recommendation*, 22:2004-3.

Liu, X., Bollen, J., Nelson, M.L. and Van de Sompel, H. (2005), "Co-authorship networks in the digital library research community", *Information Processing and Management: An International Journal*, Vol. 41 No. 6, pp. 1462-80.

Marsden, P.V. (2002), "Egocentric and sociocentric measures of network centrality", *Social Networks*, Vol. 24 No. 4, pp. 407-22.

Mika, P. and Gangemi, A. (2004), "Descriptions of social relations", paper presented at 1st International Workshop on FOAF, Social Networks and the Semantic Web, Galway.

Mizzaro, S. (2003), "Quality control in scholarly publishing: a new proposal", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 11, pp. 989-1005.

Moed, H.F. (2002), "The impact-factors debate: the ISI's uses and limits", *Nature*, Vol. 415 No. 6873, pp. 731-2.

Newman, M.E.J. (2001), "The structure of scientific collaboration networks", *Proceedings of the National Academy of Sciences*, Vol. 98 No. 2, pp. 404-9.

Odlyzko, A. (2002), "The rapid evolution of scholarly communication", *Learned Publishing*, Vol. 15 No. 1, pp. 7-19.

Resh, V.H. (1998), "Science and communication: an author/editor/user's perspective on the transition from paper to electronic publishing", *Issues in Science and Technology Librarianship*, Vol. 19, pp. 1092-206.

Scott, J. (2000), *Social Network Analysis: A Handbook*, 2nd ed., Sage, London.

Seglen, P.O. (1997), "Why the impact factor of journals should not be used for evaluating research", *British Medical Journal (Clinical Research Ed.)*, Vol. 314, pp. 498-502.

Wasserman, S., Faust, K. and Iacobucci, D. (1994), *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*, Cambridge University Press, Cambridge.

Further reading

Giles, C.L., Bollacker, K.D. and Lawrence, S. (1998), "CiteSeer: an automatic citation indexing system", *Proceedings of the 3rd ACM Conference on Digital Libraries, Pittsburgh, Pennsylvania, 1998*, pp. 89-98.

About the authors

Nikolaos Korfiatis holds a PhD Fellowship in the Department of Informatics at the Copenhagen Business School (CBS) in Denmark. He received his BSc in Information Systems from the Athens University of Economics and Business and his diploma in Engineering of Interactive Systems with focus on Information Retrieval from the Royal Institute of Technology (KTH) in Stockholm, Sweden. His research interests span the fields of Information Science and Economic Sociology with a particular emphasis to the analysis of Online Social Networks and Evaluation Mechanisms. He is the corresponding author and can be contacted at: nk.inf@cbs.dk

Marios Poulos is an adjunct assistant professor at the Department of Archives and Library Sciences at the Ionian University in Corfu, Greece. He received his PhD degree from the University of Piraeus in the field of Biological Information Processing. His research interests are in metadata and semantics research with emphasis on intelligent methods to the extraction of semantic information from various sources

George Bokos is Professor of Information Science and Head of the Department of Archives and Library Sciences of the Ionian University in Corfu, Greece. He is also the director of this department's Information Technology Laboratory. His research interests cover several issues (e.g. library automation, metadata, etc.) concerning the application of information technologies in the library and information work.