Sharing Knowledge by Conceptual Mapping: the case of EU Thesaural Interoperability

Giacomo BARTOLONI, Enrico FRANCESCONI [1]
ITTIG-CNR, via de' Barucci 20, Florence (Italy)

Abstract. The availability of public administration semantic Web services is strictly linked to the availability of knowledge resources. In particular their interoperability allows knowledge sharing and reuse, so to provide integrated services in a distributed environment. In this paper a machine learning technique for guaranteeing thesaural interoperability by conceptual mapping within an information retrieval framework is presented. In particular the case of thesaural interopearbility for cross-collection legal information retrieval services at EU level is shown.

Keywords. Thesaurus Mapping, Information Retrieval, SKOS, Support Vector Machine

1. Introduction

Legal information accessibility and sharing in the pan-European environment is essential to ensure democratic transparency, the equality of citizens' rights, but also to guarantee legal certainty.

In the legal professions, in business, as well as in the academic and public administration domains, there is a strong need of a correct understanding of legal concepts preserving their correct sense and legal value; in particular it is important to be aware of the notions of international and EU law, combined with the correct comprehension of their affinities and differences in national laws.

This need is felt both in monolingual as well as in multilingual and multicultural domains, complicated by the variety of legal terminologies in different legal systems and languages. In public administration, in particular, this need is emerging for the increasing demand of integrated semantic access services to heterogeneous legal data sources in a distributed environment: in these cases the variety of legal terminology in different legal systems and data sources may affect information retrieval quality.

While retrieval quality from a single data source is usually guaranteed by the use of specific thesaurus, a similar quality in cross-collection retrieval services can

[1] E. Francesconi is author of Sections 1, 3, 4, 5, 6, 7 and 9; G. Bartoloni is author of Sections 2 and 8.

be guaranteed by providing interoperability between different thesauri. In this context interoperability means using a particular thesaurus for users' queries and mapping it to thesauri in other languages, to more specialized vocabularies, or to different versions of the same thesaurus [1], in order to obtain a retrieval from different digital collections which is coherent to the original query.

In [2] a methodological framework for semantic mapping between thesauri, based on information retrieval techniques, was presented, as well as a specific approach within such framework on a case study aimed at mapping different thesauri of interest for the European Union institutions.

In this paper a machine learning approach able to provide training facilities to the mapping techniques within the same framework is introduced and tested. In particular in Section 2 standards and criteria used for representing thesauri are described; in Section 3 our characterization of the thesaurus mapping problem, having only schema information available, is briefly recalled; in Section 4 the approach to represent thesaural concepts is reported; in Sections 5 and 6 the functions to measure the similarity between concepts, as well as the conceptual mapping prediction approach based on Support Vector Machine (SVM), are respectively shown; in Section 7 the implementation of the proposed approach on a thesaural mapping case-study, which is of interest for the Publication Office of the European Union is presented; in Section 8 the experiments on such thesaural mapping case-study are reported and, finally, in Section 9 some conclusions are discussed.

2. Standard representation of thesauri concepts and relations

ISO has defined two international standards[2] which are useful to ensure consistency in the development of mono/multilingual thesauri within or between indexing agencies. ISO standards provide guidelines for concepts and relations but do not provide guidelines for adopting specific thesaurus data formats. In order to manage, process and compare different thesauri structures, as well as to share them in a machine readable way, the use of a common standard, able to keep the semantics of their native data formats, is essential.

The Semantic Web community has developed the SKOS[3] standard which uses RDF to represent different knowledge organization systems such as thesauri, classification schemes, subject heading systems and taxonomies, as well as to share them in a distributed environment. Following SKOS recommendations[4], a knowledge organization system can be viewed as a concept scheme including a set of concepts. The SKOS vocabulary deems a concept (identified by the `skos:Concept` class) as the most elementary unit. A concept can be connected with any number of strings, in any natural language, but with only one preferred label (in every language) while it can have infinite alternative descriptions. By the use of the `skos:prefLabel` and `skos:altLabel` properties, the preferred and the alterna-

[2]ISO5964/ISO2788 *Guidelines for the establishment and development of multilingual/monolingual thesauri*

[3]Simple Knowledge Organization System (http://www.w3.org/2004/02/skos/)

[4]developed from the 2005's to the 2009's versions.

tive descriptions are tied to the concepts. Furthermore, one or more notations (`skos:notation`, including a string of characters in any natural language) can be assigned to a SKOS concept in order to identify it in the application field of another concept scheme.

While SKOS provides a standard way to represent thesauri descriptors and relationships, in literature different approaches to represent thesauri have been proposed ([3], [4], [5]), but, so far, no standardized architecture for translating thesauri from their proprietary format has emerged. Therefore, in this case study, a knowledge architecture for representing thesauri using SKOS is proposed on the basis of similar experiences reported in literature.

2.1. A knowledge architecture to represent thesauri using SKOS

[5] is an interesting work which proposes an architecture to represent thesaural concepts and relations: it describes a structured method to convert thesauri to SKOS, evaluating the applicability of SKOS meta model to represent existing thesauri.

Starting from the method given in [5], as well as SKOS specifications, in our case-study a methodology for thesauri conversion to SKOS is proposed. In particular the following criteria have been followed: thesauri descriptor labels are represented by `skos:prefLabel`; *used-for* relations are represented by `skos:altLabel`, and different kind of *notes* are mapped to the correspondent `skos:scopeNote` and `skos:editorialNote` elements. *Broader*, *narrower* and *related* relations are directly mapped to the corresponding SKOS properties. Moreover, the native multilingual tools provided by SKOS made it easy to handle the multilingual labels connected to the concepts.

Structural patterns which haven't a direct counterpart in SKOS are represented providing extensions according to SKOS specifications [6]. For example, usually and in particular in our case-study, thesauri have a conceptual structure, organized in hierarchical levels. Therefore, in order to describe their native semantics, the `skos:Concept` class has been extended into to 3 additional classes: `eu:Descriptor`, `eu:Microthesaurus` and `eu:Domain` where '`eu`' is the namespace defined in this work for the Publication Office of the European Union thesauri SKOS extension (Fig. 1).

3. Formal characterization of Thesaurus Mapping

Thesaurus mapping for the case-study is a problem of descriptors alignment, having only thesaural schema available (*Schema-based mapping* [7]). In this case thesaurus mapping is the problem of identifying the conceptual/semantic similarity between a descriptor (represented by a simple or complex term[5]) in a source thesaurus and candidate descriptors in a target thesaurus.

These characteristics allow us to propose a characterization of the schema-based Thesaurus Mapping ($\mathcal{TM}$) problem as a problem of Information Retrieval ($\mathcal{IR}$): the aim is to find concepts in target thesaurus, better matching the seman-

[5]for example *Parliament* is a simple term, *President of the Republic* is a complex term.

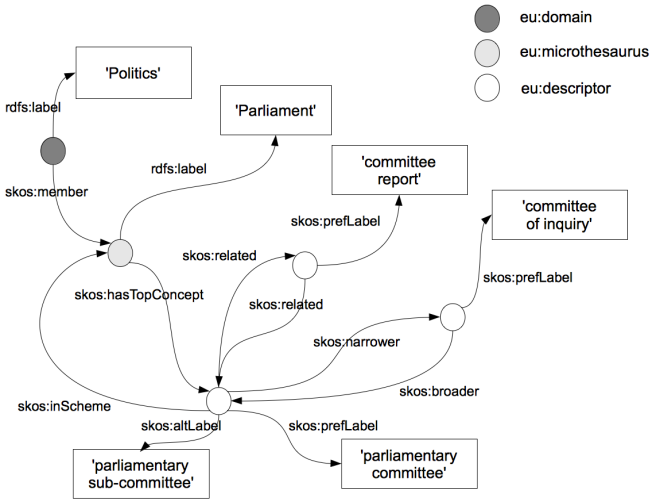

Figure 1. SKOS representation of an EUROVOC excerpt.

tics of a concept in a source thesaurus. The isomorphism between $\mathcal{TM}$ and $\mathcal{IR}$ ($\mathcal{TM} \equiv \mathcal{IR}$) can be established once we consider a source concept as a *query* of the $\mathcal{IR}$ problem, and a target concept as a *document* of the $\mathcal{IR}$ problem.

Therefore, the $\mathcal{TM}$ problem can be viewed and formalized, like the $\mathcal{IR}$ problem, as a 4-uple $\mathcal{TM} = [D, Q, F, R(\vec{q}, \vec{d})]$ [8] where:

1. D is the set possible representations (*logical views*) of a concept in a target thesaurus (a document to be retrieved in the $\mathcal{IR}$ problem);
2. Q is the set of the possible representations (*logical views*) of a concept in a source thesaurus (a query in the $\mathcal{IR}$ problem);
3. F is the framework of concepts representation in source and target thesauri;
4. $R(\vec{q}, \vec{d})$ is a ranking function, which associates a real number with $(\vec{q}, \vec{d})$ where $\vec{q} \in Q$, $\vec{d} \in D$, giving an order of relevance to the concepts in a target thesaurus with respect to a concept of a source thesaurus.

In this framework the implementation of a thesaurus mapping procedure is represented by the instantiation of the previous 4 components.

4. Logical views (Q and D) of descriptors and matching framework (F)

Mapping between thesaural concepts is a process which aims at matching concept semantics rather than their lexical equivalences. In traditional thesauri *descriptors* and *non-descriptors* are represented by different terms (`skos:prefLabel` and `skos:altLabel`, according to SKOS) expressing the same meaning. More precisely, each meaning is expressed by one or more terms[6] in the same language (for instance 'pollution', 'contamination', 'discharge of pollutants'), as well as in

[6]Linguistic expressions by single or multi words.

different languages (for instance, the English term 'water' and the Italian term 'acqua', etc.). Moreover each term can have more than one sense, i.e. it can express more than one concept. Therefore to effectively map thesaural concepts, term (simple or complex) semantics has to be captured and represented.

In $\mathcal{IR}$ a query is usually constructed as a context (set of keywords) able to better represent the semantics of a query. Similarly in $\mathcal{TM}$ the semantics of a thesaural concept is conveyed not only by its terms, but also by the context in which the concept is used as well as by the relations with other concepts. In $\mathcal{TM}$ problem, Q, D and F are exactly aimed at identifying logical views and related framework for concept representations able to better capture the semantics of terms in source and target thesauri, as well as to measure their conceptual similarity.

In this work we propose to represent the semantics of a thesaural concept by a vector $\vec{d}$ of binary[7] entries composed by the term itself, relevant terms in its definition, in the alternative labels, as well as terms of directly related thesaural concepts (broader, narrower, related concepts).

Firstly a vocabulary of normalized terms from target thesaurus is constructed, where 'normalization' in this context means string pre-processing, in particular word stemming and stopwords eliminations. Being T the dimension of such vocabulary, both source and target concepts $\vec{d}$ are represented in a vector space of T-dimension ($\vec{d} = [x_1, x_2, \cdots, x_T]$); the entry x_i gives information on the presence/absence of the corresponding i^{th} vocabulary term among the terms characterizing the concept $\vec{d}$. In Fig. 2 a binary vector representation of a EUROVOC concept is sketched. In such representation the framework F is composed of T-dimensional vectorial space and linear algebra operations on vectors.

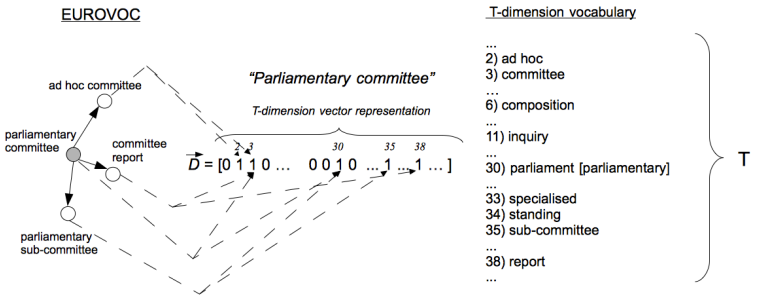

Figure 2. T-dimension vectorial representation of a thesaural descriptor $\vec{d}$.

5. The proposed ranking function (R)

Having represented the semantics of thesaural concepts as a binary vector, their similarity can be measured as the related binary vectors correlation, quantified, for instance, as the cosine of the angle between them

[7]Statistics on terms to obtain weighted entries are not possible since document collections are not available (*schema-based thesaurus mapping*)

$$sim(\vec{q}, \vec{d}) = \frac{\vec{q} \times \vec{d}}{|\vec{q}| \cdot |\vec{d}|} \tag{1}$$

where $|\vec{q}|$ and $|\vec{d}|$ are the norms of the vectors representing concepts in source and target thesauri, respectively.

6. A machine learning technique for conceptual mapping prediction

Having established a proper similarity measure between thesaural concepts, a criterion able to predict matching concepts has to be defined.

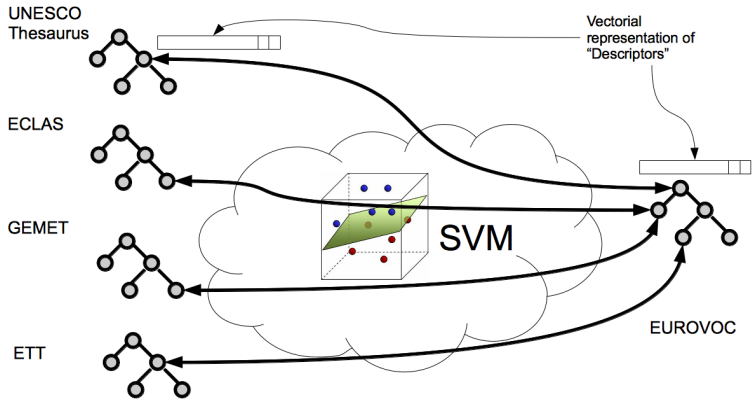

Figure 3. Thesaurus mapping using SVM

In [2] a criterion was implemented by defining a heuristic threshold over a similarity measure: if the similarity between concepts is over a threshold, a `skos:exactMatch` relation is established. Such strategy, anyway, usually suffers from generalization capabilities out of the matching examples used to tune the heuristics. Generalization capabilities for a prediction strategy can be introduced by adopting machine learning techniques able to learn a predictive function from a training set of matching relations. In this work such predictive function is obtained by a Support Vector Machine (SVM) (Fig. 3) trained to classify a pair of descriptors into two classes {Match (+1), no-Match (-1)}.

A training set for the SVM thesaurus matching predictor is composed by training examples described by vectors of features deemed representative for descriptors conceptual matching: in particular the i^{th} example is represented by a feature vector Φ_i associated to a pair $(\vec{q}, \vec{d_i})$ of a source and target thesaurus concepts respectively, including:

- the similarity measure $sim(\vec{q}, \vec{d_i})$, computed according to the cosine function (see eq. (1));
- the logical view of the target descriptor $\vec{d_i}$

together with a relevance judgment $y = \{+1, -1\}$ for that target descriptor $(\vec{d_i})$ on that source descriptor $(\vec{q})$ that is either *matching* (+1) or *non-matching con-*

cept (-1). Therefore a generic i^{th} training example describing a pair of thesaural descriptors and related relevance judgement is

$$\Phi_i = \langle \langle sim(\vec{d_i}, \vec{q}), \vec{d_i} \rangle, y_i \rangle \tag{2}$$

On the basis of such training set, the goal is to build an SVM classifier (a separating surface) which is able to distinguish between matching and non-matching descriptors. The SVM classifier provides also the distance of the examples from the separating surface, giving a measure of the prediction confidence, thus allowing a ranking among candidate target descriptors. The best ranked descriptor is finally chosen as the predicted matching concept.

The training set for the case-study is constructed on the basis of a "gold standard" matching concepts data set, built by human experts from a set of thesauri of interest for the European institutions.

7. Case-study interoperability assessment through a "gold standard"

In this work a thesaurus interoperability case-study is proposed, including five thesauri of interest for the Publication Office of the European Union. The thesauri are EUROVOC, ECLAS, GEMET, UNESCO Thesaurus and ETT.

EUROVOC is the main EU thesaurus containing a hierarchical structure with inter-lingual relations. It helps to coherently and effectively manage, index, and search information of EU documentary collections, covering 21 fields. ECLAS is the European Commission Central Libraries thesaurus, covering 19 domains. GEMET, the GEneral Multilingual Environmental Thesaurus, is utilised by the European Environment Agency (EEA). UNESCO Thesaurus is a controlled vocabulary developed by the United Nations Educational, Scientific and Cultural Organisation which includes subject terms for several areas of knowledge. ETT is the European Training Thesaurus providing support for indexing and retrieval vocational education and training documentation in the European Union.

As introduced in Section 6, interoperability between thesauri has been assessed on a "gold standard" data set, namely an ideal collection of conceptual mappings expected by humans. To build the "gold standard" data set, an intellectual activity has been carried out by two groups of experts, dealing with EUROVOC as pivot thesaurus on the "Law" or "Employment and Working conditions" domains, chosen to assess the interoperability approach. The experts have established exact match relations between EUROVOC descriptors and the descriptors of the other thesauri. Specific guidelines have been given to the experts [9] to establish mapping relations, limited to the `skos:exactMatch` relation. Using a tool (THALEN[8]) [2] developed within the project, the experts are able to establish `skos:exactMatch` relations.

The work of legal experts, in their commitment to reach a reliable matching among thesauri, has been harmonized through several meetings, in order to identify common criteria to build the "gold standard". Such meetings raised a number of critical considerations about the activities of thesaurus mapping as carried

[8] THesaurus ALigning ENvironment

out by experts, as well as they gave the feeling of the complexity of the task to be carried out by machines. In Tab. 1 some paradigmatic cases of the criteria adopted by experts to establish mapping relations are shown and discussed.

Eurovoc `skos:prefLabel`	**ETT** `skos:prefLabel`	**Notes**
craftsman	craftsman	exact string matching
dismissal	termination of employment	use of synonyms
holding of two jobs	multiple employment	target descriptor definition (*Where a worker holds more than one job at the same time, legally, either for two or more different employers or as self-employed for one of the jobs.*)
long-term unemployment	long term unemployment	same terms, even if with different morphological manifestation
non-standard employment	non traditional occupation	use of expert background knowledge

Table 1. Paradigmatic human criteria adopted for establishing thesaurus mapping relations (EUROVOC-ETT case).

A part from trivial cases represented by pure string similarity, from this short survey some conclusions can be derived. Basically the criteria used by the experts followed mental deductions which derived by a deep semantic analysis of the information associated to descriptors, regarding related terms as well as the typologies of the relations, definitions, analysis of inheritance properties, pure human background knowledge.

This activity produced the "gold standard" data set reported in Tab. 2. The available versions of EUROVOC, UNESCO Thesaurus and ETT, as well as the related gold standards, are characterized by a well organized structures, including preferred and alternative labels as well as thesaural relations between descriptors. On the other hand the available versions of ECLAS and GEMET are less structured, the result is an ECLAS gold standard characterized by few conceptual relations and a GEMET gold standard characterized by few relations and alternative labels. The experiments therefore have been carried out only on the cases able to provide meaningful statistics, as reported in Section 8.

Thesauri	`skos:exactMatch` **relations**
EUROVOC-ETT	131
EUROVOC-UNESCO	93
EUROVOC-ECLAS	143
EUROVOC-GEMET	28
Total exact match	395

Table 2. The "gold standard" of exact matching concepts.

8. Experiments

A set of experiments on the SVM model for thesaural conceptual mapping is conducted over the "gold standard" data set, which is used to build examples for SVM training and test. The SVM training set includes both "gold standard" matching descriptors, as well as an equal number of non-matching descriptors in order to balance the training set and to allow the system to distinguish between matching and non-matching concepts.

To measure the SVM classifier[9] performances, a *k-fold* cross-validation strategy has been developed. The examples have been divided into $k = 3$ groups: 2 of these groups are, alternatively, used to train the classifier while the remaining group is used to test the system. The classification accuracy is computed as the fraction of correct tests over the entire number of tests. Tabs. 3, 4, 5, 6 report *k-fold* cross-validation accuracy, which is computed as the average accuracy over the $k = 3$ runs. Descriptors are represented using terms contained in the `skos:prefLabel`, moreover different combinations of information coming from either `skos:altLabel` or obtained in related descriptors, if any, are used.

altLabel	Related concepts	Accuracy
no	no	83,87%
yes	no	93,55%
no	yes	100%
yes	yes	100%

Table 3. EUROVOC-UNESCO mapping

altLabel	Related concepts	Accuracy
no	no	87,02%
yes	no	95,42%
no	yes	100%
yes	yes	100%

Table 4. EUROVOC-ETT mapping

altLabel	Related concepts	Accuracy
no	no	93,00%
yes	no	93,71%

Table 5. EUROVOC-ECLAS mapping

altLabel	Related concepts	Accuracy
no	no	100,00%

Table 6. EUROVOC-GEMET mapping

In the EUROVOC vs. UNESCO (Tab. 3) and ETT (Tab. 4) experiments better results have been obtained using information from `skos:prefLabels`, `skos:altLabel` and related concepts, rather than from `skos:prefLabel` only, or from `skos:prefLabel` and `skos:altLabel` only, so to confirm the validity of the approach. Similarly the EUROVOC vs. ECLAS mapping (Tab. 5) reached better results using information from both `skos:prefLabel` and `skos:altLabel`, rather than from `skos:prefLabel` only, while no meaningful statistics can be obtained by using information from related concepts. On the other hand for EUROVOC vs. GEMET experiments (Tab. 6) only statistics related to the use of `skos:prefLabel` can be given, which nevertheless showed very good performances.

9. Conclusions

The development of thesaurus mapping services is considered a relevant issue for the complete implementation of eGovernment policies, as stated in [10], in order to

[9] http://www.cs.cornell.edu/People/tj/svm_light/

provide semantic and cultural interoperability of public services and information quality. Interoperability among thesauri, on the one hand, contributes to overcome cultural and language differences which jeopardise effective communication and action across different countries and governmental bodies; on the other hand, it meets the need of governments, markets and individuals to have accurate and appropriate information from different sources.

In this work a machine learning methodology within a specific framework for thesaurus mapping, having only schema information available, has been implemented and tested. The approach has been assessed on a case-study focused on five thesauri of interest for the EU institutions.

Two main problems have been addressed: how to represent the semantics of thesaural concepts, and how to provide effective tools to implement an automatic mapping between them, to be validated by human experts. While semantics of thesaural concepts has been represented in a vectorial space, an SVM approach has been trained and used to provide matching prediction over a similarity measure between vectors.

The experimental results give evidence of the reliability of the approach, outperforming, on a wider data set, the results obtained in [2]. Further experiments can be foreseen by using different similarity measures within the same identified $\mathcal{TM}$ framework, within which different $\mathcal{IR}$ techniques can be implemented, thus facilitating tuning and cross-validation of different $\mathcal{TM}$ approaches.

References

[1] M. Doerr, "Semantic problems of thesaurus mapping," *Journal of Digital Information*, vol. 1, no. 8, 2001.

[2] E. Francesconi, S. Faro, and E. Marinai, "Thesauri alignment for eu egovernment services: a methodological framework," in *Proceedings of the JURIX 2008 Conference*, pp. 73–77, IOS Press, 2008.

[3] J. Neubert, "Bringing the "thesaurus for economics" on to the web of linked data," in *Proceedings of the WWW2009 Workshop on Linked Data on the Web* (T. B.-L. K. I. Christian Bizer, Tom Heath, ed.), vol. 538 of *CEUR Workshop Proceedings*, CEUR-WS, April 2009.

[4] A. Isaac, S. Wang, C. Zinn, H. Matthezing, L. van der Meij, and S. Schlobach, "Evaluating thesaurus alignments for semantic interoperability in the library domain," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 76–86, 2009.

[5] M. V. Assem, V. Malais, A. Miles, and G. Schreiber, "A method to convert thesauri to skos," in *Volume 4011 of Lecture Notes in Computer Science*, pp. 95–109, Springer, 2006.

[6] A. Miles and S. Bechhofer, "Skos simple knowledge organization system reference," in *http://www.w3.org/TR/skos-reference*, W3C Semantic Web Deployment Working Group, 2009.

[7] E. Rahm and P. Bernstein, "A survey of approaches to automatic schema matching," *The International Journal on Very Large Data Bases*, vol. 10, no. 4, pp. 334–350, 2001.

[8] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.

[9] A. C. Liang and M. Sini, "Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures," *New Review of Hypermedia and Multimedia*, vol. 12, no. 1, pp. 51–62, 2006.

[10] X. Ma, M. A. Wimmer, S. Dawes, M. Bicking, C. Codagnone, and M. Janssen, "eGovernment R&D Roadmap 2015," in *Expanding the Knowledge Economy: Issues, Applications, Case Studies* (P. Cunningham and M. Cunningham, eds.), IOS Press, 2007.