MAPPING THE W78 PAPERS ONTO THE CONSTRUCTION INFORMATICS TOPIC MAP

Dr. Žiga Turk, Assoc. Prof.,
University of Ljubljana, Slovenia
ztурk@itc.fgg.uni-lj.si

Dr. Tomo Cerovšek, Asst.
University of Ljubljana, Slovenia
tcerovse@fgg.uni-lj.si

SUMMARY

In the context of the EU ICCI project, an ontology of the filed of construction informatics (alias construction information technology) has been developed and defined in the form of an ISO/IEC 13250 topic map. The map is based on a generic model of a research process. It structures the field into core and support themes, the core are further split into information and communication related topics. In the context of the EU SciX project a full text electronic bibliography of the CIB W78 since 1988 has been compiled. In the paper we present the results of the mapping of the W78 papers onto the topic map. Two approaches were taken: in the first the papers were mapped into the topics used some heuristics and human intervention. We present statistical analysis and chronology of the topics. In the second, data mining methods were used to define the topic map of W78 automatically. The resulting structure of topics is based purely on the words and phrases of the papers and not on some higher-level structure. We compare this machine generated map with the topic map invented by a human. The latter resulted in topics of more similar sizes and sometimes curious interdependencies.

KEYWORDS: Construction informatics, information technology in construction, domain definition, bibliography.

INTRODUCTION

Since 1982 CIB W78 has been providing a cradle to what has today emerged as a scientific field within construction and what we call construction informatics, construction information (and communication) technologies. Until recently, it has been defied more or less informally. Some recent works, however, attempted a more formal definition. In this paper we restate the definition of the field and present a structured set of topics (Chapter 2). These topics are then used to analyse the papers that have been published in this community since 1988. Chapter 3 presents some general statistics and Chapter 4 an attempt to map the papers onto the topic map.

The "academic" motivation of the work presented in this particular paper is to (1) perform introspective analysis of what has been going on in the community in the past 15 years, to see (2) if an "objective" computer made understanding of the topics would result in similar categories as did a human intellectual effort and (3) to explore some machine learning techniques, that can later be reused for a more general knowledge management in construction.

The practical motivation is to create a navigation structure across the digital library of W78 with a minimum of human intervention.

TOPIC MAP OF CONSTRUCTION INFORMATICS

Ontology is one of the four attributes (in addition to axiology, methodology and epistemology) that define a field of science. Ontologies may be represented in various ways, one is a topic map for which an ISO standard is available - ISO ISO/IEC 13250 Topic Maps. This standard "*provides a standardized notation for interchangeably representing information about the structure of information resources used to define topics,*

and the relationships between topics." In the ICCI project such a topic map of construction informatics was defined (Turk et al., 2002). In this paper we only present the main points.

Related work

One of the firsts attempts, to literary map this field, is a rising coastline drawing from VTT (Hannus, 1996). It provided a metaphor for one issue that construction informatics is dealing with - computer integrated construction. Fenves (1996) defines the topic by studying its historical development. Brandon and Betts (1997) introduced a technology-oriented view on IT in construction and divide the domain into four fields: visualization, intelligence, communication, and integration. Probably the first formal definition of the field is the INFOMATE model (Bjoerk, 1996 and 1999). Turk (1997) evolved this idea and, based on the work of Winograd and Flores (1987) introduced the distinction between processing activities and commitment-negotiation or communication activities. Studying research themes both in Sweden and abroad as well as taking into account theoretical ontological backgrounds, Ekholm (2002) identified these main construction informatics themes: process and product models, classification and standardization, software applications, communication and information environments, work organization and processes and IT-strategies. In their overview of construction IT research, Amor et al. (2002), identified major themes (computer integrated construction, construction process and decision support) and technical themes (process modeling, product modeling and documents). Most of these efforts provide little argument as to why these and not some other categories were selected and do not explicit an underlying model or schema.

Underlying model

The presented topic map is based on a generic process model of research (Figure 1). Industry needs and general human curiosity generate problems and questions. The research process, done by academia and other researchers, using the knowledge from other disciplines and principles of scientific investigation, results in knowledge (as well as new questions). The knowledge is then used in (1) teaching, (2) development of technology and (3) development of standards, codes, best practices etc. Through these three main mechanisms this new knowledge reaches the industry and affects the day to day business processes.

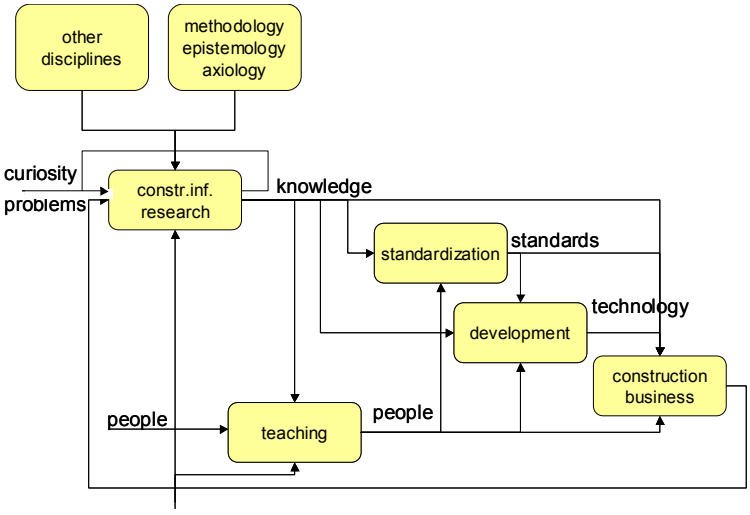

Figure 1 Simplified IDEF0-ish process model of construction informatics.

The top level classification of construction informatics themes is therefore into:

- core themes that create the knowledge.
- support themes, in which the need is identified, knowledge transferred, impacts measured etc.

The above breakup is the current proposal for the top level construction informatics themes:

- strategies,

- core research,
- transfer,
- impact.

To structure the core themes, some kind of a generic model of construction industry is required. Should technology assist humans we distinguish between two kinds of human work. Communication (commitment negotiation works) and processing works. Communication is further split by the actors involved. Processing is split according to information life-cycle: it is first created, then represented and formatted, then stored and finally retrieved. Detailed explanation is given in the ICCI D212 report (Turk at al., 2002).

Hierarchy of construction informatics topics

The classification of the core research topics that will be used in this paper is given below. Associated keywords are given in brackets.

Core themes:

- communication (communication, discussion, collaboration)
  - man-man (email, internet, mobile, chat, conference, groupware, workflow)
  - man-software (user interface, interaction, windows, mouse)
  - software-software (software, program, data exchange, corba, soap, xml, com, dcom)
  - software-machine (software, program, robot, sensor)
- processing
  - create information
    - analysis (analysis, finite elements, design, structural analysis, structural design,
    - synthesis (drafting, computer aided design, 3d modeling, cad)
  - manage information
    - represent (format, schema, ontology, data structure, model, standard, STEP, IFC, XML)
    - store (database, relational, document management)
    - retrieve (information retrieval, data mining, search, query, classification, thesaurus, vocabulary, glossary)
- common infrastructures
  - collaboration (internet, project web, web, portal, communication)
  - commerce (ecommerce, commerce, business)
  - legal (legal, standard, law, regulation)

Support themes:

- needs
  - roadmaps (needs, survey, roadmap, vision, future)
  - strategies (needs, survey, strategy, plan, business, reengineering)
- transfer
  - bestPractise (best practise, example, knowledge transfer)
  - education (education, teaching, learn, knowledge transfer
  - software development (software, program, prototype)
  - standards (standard, iso, ifc, xml, bcxml)
- deployment (experience, lesson, learn)
- impact (impact, result)
  - economic (impact, result, saving, efficient)
  - environment (impact, result, environment, clean)
  - social (social aspect, social responsibility, social consequences, community)

W78 DIGITAL LIBRARY

In the context of the SciX project (www.scix.net) an on-line bibliography of construction informatics is being compiled. Currently (March 2002) it includes about 870 papers from the ECCE, W78 conferences and the ITcon journal. We hope to include some other tracks of conferences, such as EG-SEA-AI, ECPPM, CE and

journals as well as personal and institutional bibliographies from the field. A policy has been suggested only to include works where the full text is available for free. Table below summarizes the number of papers collected so far. Some were available in digital form from the beginning, some had to be digitised from paper copies.

Statistics

The number of W78 papers in the library is 689. On each paper we have the metadata and the full text in the PDF format. The metadata is structured into these fields: authors, title, summary, source, pages, keywords (where available), series (e.g. w78:2002) and class (to be discussed later). The metadata can be exported in a Dublin Core compatible format. The size of the pdf files ranges from 4 kilobytes to 4 megabytes and totals half a gigabyte.

Table 1 Number of papers in the library. 1999 (Vancouver) only includes the papers from the 4th volume of the proceedings. The 2000 (Iceland) was a joint workshop by IABSE WC6 and EG-SEA-AI and all papers are included.

1. w78:1988 (32)
2. w78:1991 (26)
3. w78:1992 (41)
4. w78:1993 (83)
5. w78:1994 (38)
6. w78:1995 (55)
7. w78:1996 (47)
8. w78:1997 (42)
9. w78:1998 (44)
10. w78:1999 (76)
11. w78:2000 (101)
12. w78:2001 (42)
13. w78:2002 (62)

Table 2 Most frequent authors, words and phrases.

Authors	Frequent words in titles and summaries	Most common 2 word phrases
Tolman (15)	construct	construct;industri
Froese (13)	design	product;model
Turk (13)	model	inform;technologi
Amor (12)	inform	paper;describ
Fischer (12)	build	construct;process
Vanier (11)	project	inform;system
Dawood (10)	system	life;cycl
Alshawi (9)	process	construct;project
Aouad (9)	develop	project;manag
Rezgui (9)	paper	paper;present
Bjork (8)	manag	build;design
Cooper (8)	integr	knowledg;base
Fridqvist (8)	data	object;orient
Howard (8)	product	design;process
Bouchlaghem (7)	industri	design;construct
Christiansson (7)	base	project;inform
Eastman (7)	present	data;model
Ekholm (7)	applic	build;construct
Hannus (7)	comput	data;exchang
Rebolj (7)	knowledg	process;model
Sariyildiz (7)	technologi	product;data

Evolution of topics

The figure below shows the relative popularity of some keywords and phrases as present in the titles and abstracts. This can be compared with a similar figure from a paper by Martens and Turk (1999) that presented a study for the community of computer aided architectural design and the CUMINCAD database (http://cumincad.scix.net/).

MAPPING THE W78 PAPERS INTO THE TOPIC MAP

The topic map presented in Section 2 is a personal, subjective view on construction informatics. In this section we attempt to validate the map in two ways:

- we map the papers from W78 into the topic map by using the keywords by which the topics in the map are described.
- we use AI techniques to create clusters (groups) of papers independently and then compare the results with the first two methods.

Topic map can be considered validated if (1) any paper can be categorised with a high enough confidence and (2) if the overlap between categories is small.

Related work

The work presented in this and the previous sections is based on the text mining, statistical text analysis methods and information retrieval. In the field of construction and architecture, the Maher and Simoff (1998) reported on using these technologies in project data management. Martens and Turk (1999) were using these techniques to analyse the field of computer aided architectural design. That paper also includes bibliography on the topic of statistical text analysis.

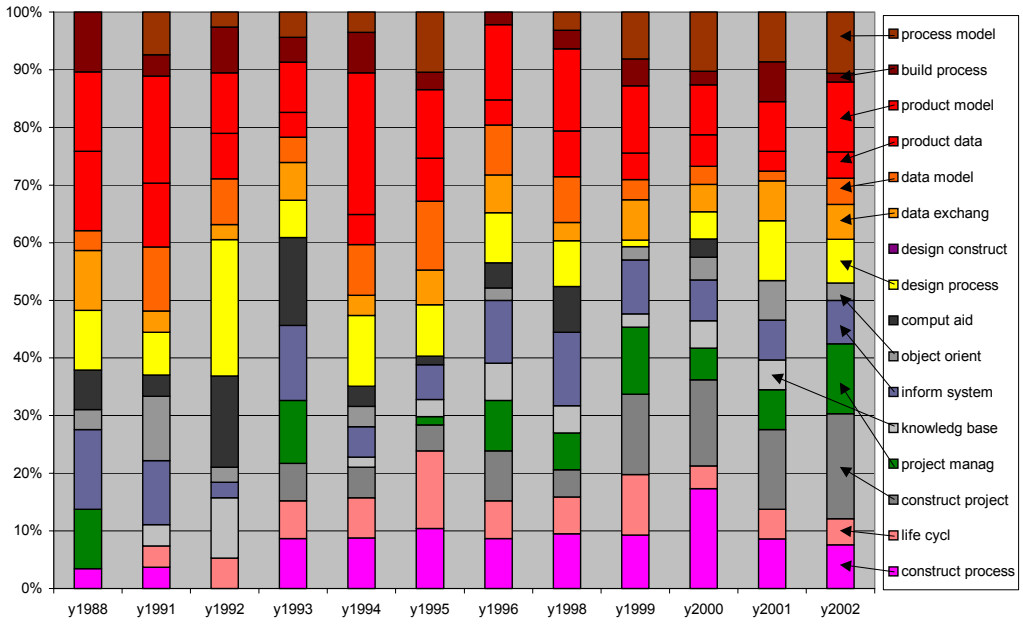

Figure 2 Relative frequency of some phrases.

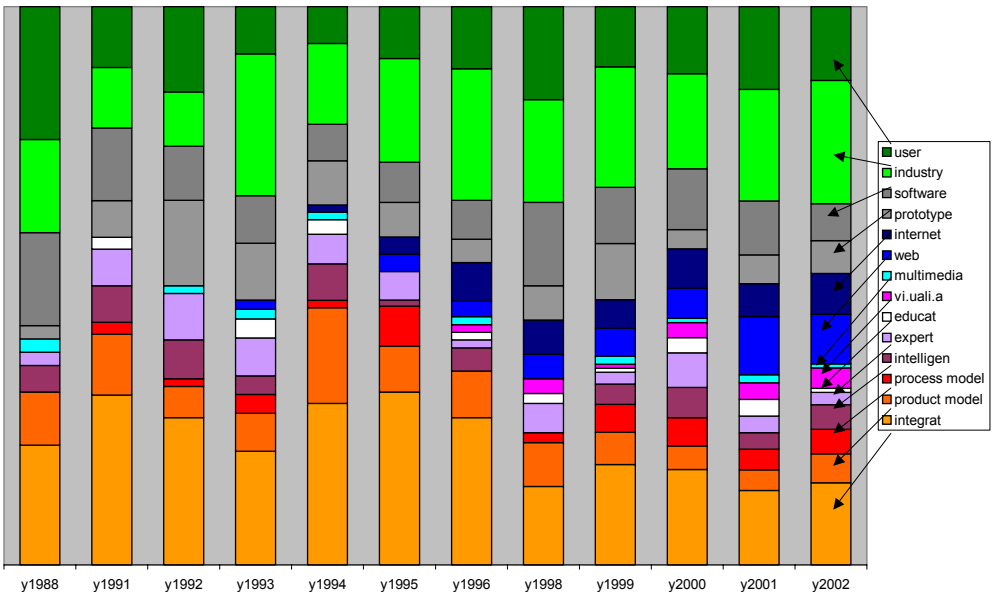

Figure 3 Relative frequency of some words.

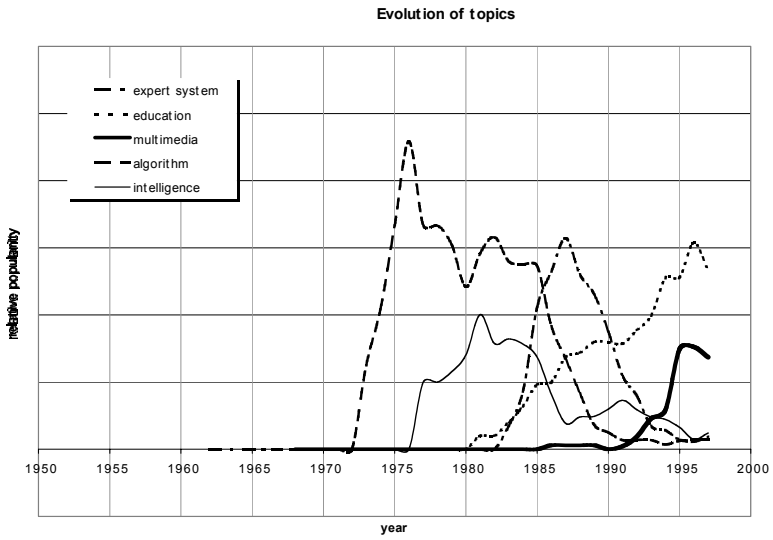

Figure 4 Evolution of themes in CAAD.

Mapping the W78 onto the topic map using keywords

Cosine similarities between the papers and the keywords in the map were computed. The table below presents the number of papers fitting each of the categories (as defined in 2.3) with a given cosine similarity. Similarity of 200 is a very good similarity, similarity of 10 is a very poor one. We can see that some categories are very broad. A study of the overlap between the categories will show the quality of the classification. Initial results show that the categories in the "processing" branch may not be well computable from the word statistics.

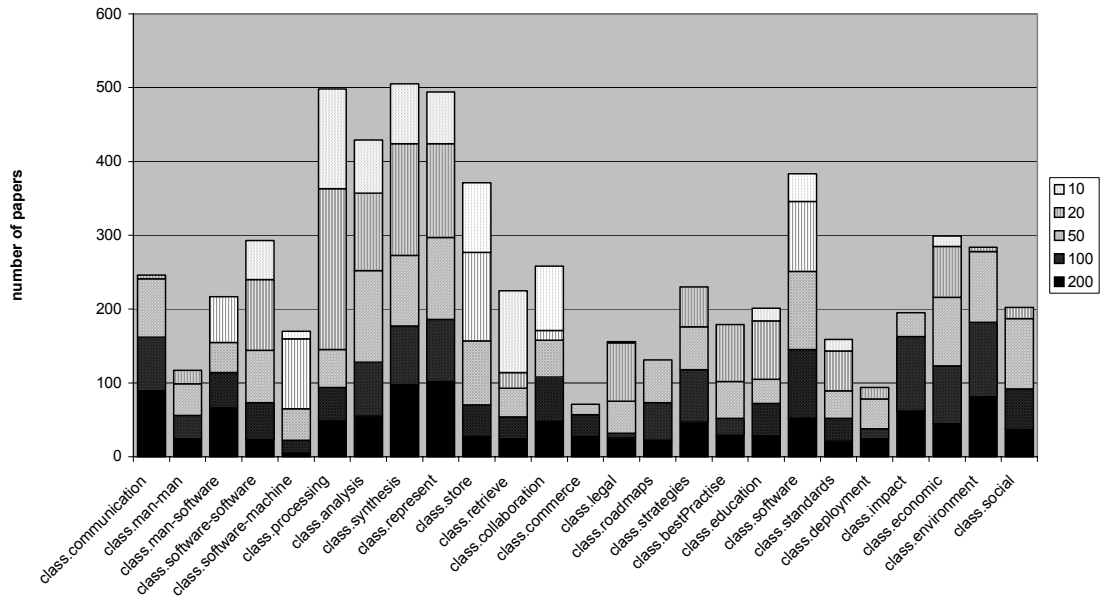

Figure 5 Confidence of classification by keywords

Machine generated topic map

Research on statistical text analysis also resulted in methods that can create topic maps automatically. To put it more precisely, based on the statistics of words and phrases in a paper, clusters of similar papers may be identified. This can be done in such a way that these clusters are organised into a hierarchy. The available algorithms include top down and bottom up clustering. We did a top down clustering of the papers in the w78 digital library and asked for depth of three (like a directory structure going three levels deep) and branching factor of three (the library is broken into three parts, each of the three parts into another three parts etc.).

The figure below outlines the topics and their most important keywords, as identified by the software. The results may be interpreted as if in the 0 branch there are papers dealing with problems from the industrial/management perspective, branch 1 includes technology oriented topics and branch 2 more theoretical work related to modelling. This interpretation is, of course, subjective. The reader is welcome to make his own conclusions, based on words that were found significant by the software.

/0/	project inform construct industri system applic manag data technologi paper
/0/0/	osmo servic condor access user includ infrastructur enterpris base ve
/0/1/	construct project industri process manag evalu system construct;industri survei improv
/0/2/	knowledg construct design build learn organis inform industri process architectur
/1/	inform technologi effect product analysi environ human build position effect;inform
/1/0/	process project design cobrit client improv support aim propos client;design
/1/1/	manag inform;technologi technologi percept asset;manag asset staff studi support manag;support
/1/2/	servic inform industri connet internet construct construct;industri centr nation site
/2/	model design build system product data process develop inform construct
/2/0/	design model build process inform product project system integr construct
/2/1/	data model standard product exchang inform applic develop project integr
/2/2/	design build system model simul process develop tool comput data

Figure 6 Construction informatics topics and associated keywords - machine made. Keywords are stemed.

Man-made map

Ideally, one should sit down and assign each paper a category and in this way put the W78 digital library in order. This is quite a lot of work, besides, it would be subjective. However, by observing user behaviour at the Website such maps may be created automatically. Since we do not have the data now, this is a theme for future work.

CONCLUSIONS AND DISCUSSION

The paper presented an initial investigation into the field of construction informatics and attempted an objective analysis of the field based on statistical text analysis of the titles and abstracts. The topic map initially proposed was found useful, however, there was a substantial overlap and difficulty to assign quite a few papers to a certain category with a high enough level of confidence using statistical text analysis - same words are used in the abstracts of the papers that should be in different categories. On the other hand, machine made classification suggested some other topics to be used. The proposed intellectually made topic map therefore seems unpractical because it would require an expert to manually classify the papers into those categories. Before we give up, however, a few other techniques will be tried, for example manually categorising just a few papers and see if the software can "learn" from examples or by learning from the user behaviour. Results of this work are immediately visible on the Web at http://itc.scix.net/.

ACKNOWLEDGEMENTS

The presented work has been conducted in the context of the SciX project, funded by the European commission under the contract IST - 2001 - 33127 and the ICCI project funded by the European commission under the contract IST-2000-33022. The contribution of the funding agency and of the partners is appreciated.

The ITC Digital Library was made possible by the organisers of the numerous W78 workshops and seminars and by the effort of Mateja Šmid and Etiel Petrinja who processed the papers.

REFERENCES

Amor R, Betts M, Coetzee G, Sexton M (2002), Information Technology for Construction: Recent Work and Future Directions Electronic Journal of Information technology in Construction, Vol. 7, pg. 245-158.

Bjoerk, B-C. (1996). Information technology in construction, Proceedings of the CIB-W65 symposium, Glasgow.

Bjoerk, B-C. (1999). Information Technology in Construction: Domain Definition and Research Issues, CIDAC, Vol. 1, No.1.

Brandon, P. and Betts, M. (1997) Veni, Vidi, Vici; in Brandon P and Betts M (editors), The Armathwaite Initiative, ISBN 1-900491-03-6, University of Salford, UK.

Ekholm, A. (2002). Themes and projects of ”IT bygg&fastighet 2002”– an analysis as a background for further initiatives, Lund Institute of Technology/Lund University, http://www.itbof.com/sa/node.asp?node=220

Fenves, S.J. (1996). Information technologies in construction: a personal journey, in Z. Turk (editor), Construction on the information highway, ISBN 961-6167-11-1, CIB Publication 198, University of Ljubljana, Slovenia, pp. 20, http://www.fagg.uni-lj.si/bled96/

Hannus, M. (1996). Islands of Automation in Construction, in Z. Turk (editor), Construction on the information highway, ISBN 961-6167-11-1, CIB Publication 198, University of Ljubljana, Slovenia, pp. 20.

ISO/IEC FCD 13250 Topic Maps (1999). http://www.ornl.gov/sgml/sc34/doocument/0058.htm

Maher, M.L. and Simoff, S. (1998). Knowledge discovery from multimedia case libraries, in I. Smith (ed) Artificial Intelligence in Structural Engineering, Springer, Berlin, 197-213.

Martens, B. and Turk, Ž. (1999) Working Experiences with a Cumulative Index on CAD: "CUMINCAD", Proceedings, ECAADE99 Conference, Turing to 2000, Liverpool, September 15-17, 1999, pg. 327-333.

Negroponte, N. (1975). Soft architecture machines. Cambridge, Mass., MIT Press.

Ontoweb (2002). A survey on ontology tools, Deliverable 1.3 of IST-2000-29243, Ontology-based information exchange for knowledge, management and electronic commerce.

Turk Ž, Šmid M, Cerovšek T, Reflak J. (2002). ICCI ontological framework and classification, report, University of Ljubljana.

Turk, Ž. (1997). A Framework for Engineering Information Technologies, K.S. Pawar (editor), Proceedings, International conference on concurrent enterprising, University of Nottingham, 8-10 Cctober 1997, ISBN 0 951 9759 6X, pp. 257-268.

Turk, Ž. (2001). Internet Information and Communication Systems for Civil Engineering - A Review, Chapter 1 in B.H.V. Topping (ed.), Civil and Structural Engineering Computing:2001, Saxe-Coburg Publications, Scotland, ISBN 1-874672-15-6, pg.1-26.

Turk, Ž. (2002). Elements of an Ontology of Construction Informatics, Rezgui Y, Ingirige B, Aouad G (ed.), Proceedings of the European Conference on Information and Communication technology Advances and Inovation in the Knowledge Society - eSMART 2002, University of Salford, ISBN 0902896415, pg. 155-167.

Turk, Ž. and B. Lundgren (1999). Communication Workflow Perspective on Engineering Work, CIB Publication 236, VTT, Finland, pg. 347-356.

Winograd T and R Flores (1987). Understanding Computers and Cognition, Addison-Wesley.